EXTERNAL SCIENTIFIC REPORT

APPROVED: 7 November 2016

doi:10.2903/sp.efsa.2016.EN-1118

Development and testing of the media monitoring tool MedISys for the monitoring, early identification and reporting of existing and emerging plant health threats

[1]Universitat de Lleida (UdL) and [2]Institute for Food and Agricultural Research and Technology (IRTA)

Oscar Alomar[2], Assumpció Batlle[2], Josep Maria Brunetti[1], Roberto García[1], Rosa Gil[1], Toni Granollers[1], Sara Jiménez[2], Amparo Laviña[2], Carme Reverté[2], Jordi Riudavets[2], Jordi Virgili-Gomà[1]

Abstract

MedISys is a media monitoring system initially intended for news items related to human health. The tool has how been extended by the Joint Research Centre, Universitat de Lleida and IRTA to also deal with plant health threats. This EFSA-funded project was based on a knowledge representation approach that generated an ontology, a formal representation of knowledge related to plant health threats. The ontology models plant pests and diseases, together with other concepts related with them: affected crops, hosts, vectors and symptoms. First of all, a collection of news sources related to plant health threats was collected to be monitored by MedISys. These sources included already known manually curated Web pages but also additional ones discovered by performing global Web searches using terms appearing in the ontology. Then, the news items coming from these sources were filtered using MedISys using a set of categories with keywords to identify those actually about plant health threats. Most of these categories focused on known threats and used terms associated with the 117 pests and diseases selected at the beginning of the project. Additionally, categories for unknown threats were also developed. In this case the categories included keywords that are usually used by experts to describe unknown threats and keywords related with symptoms expressions. All these MedISys categories combined provide mechanism to monitor plant health threats mentions in media, from newspapers to social media, ranging from those that explicitly mention a named threat (useful to monitor re-emerging threats or their spread) to those related to unknown ones (to monitor potential new threats). The project concluded with an evaluation of the e-mail alerts and reports generated by MedISys based on the previous categories. A survey and tests with real users were conducted and the results analysed to generate a set of recommendations and improvements to facilitate the use of MedISys as a plant health threats monitoring tool.

Key words: media, monitoring, ontology, pest, plant health threat,

Question number: EFSA-Q-2013-00661

Correspondence: alpha@efsa.europa.eu

Acknowledgements: The authors would like to thank EFSA staff (Caroline Merten, Marco Pautasso, Giuseppe Stancanelli, Agnes Rortais and Sybren Vos) and JRC staff (Jens Linge) for their support during the development of this project.

Suggested citation: Alomar, O., Batlle, A., Brunetti, J.M., García, R., Gil, R., Granollers, T., Jiménez, S., Laviña, A., Reverté, C., Riudavets, J., Virgili-Gomà, J., 2016. Development and testing of the media monitoring tool MedISys for the monitoring, early identification and reporting of existing and emerging plant health threats. EFSA supporting publication 2016:EN-1118. 81 pp. doi:10.2903/sp.efsa.2016.EN-1118

ISSN: 2397-8325

Summary

The project is driven by the 4 objectives as requested in the call for tender. Objective 1 is to collate new and appropriate media and information sources (e.g. journals, magazines, webpages, etc.) at global level, to be appended to MedISys for the screening of plant health threats. Objective 2 is to develop and test approaches and strategies to monitor re-emerging plant health threats on global and regional scales (e.g. new outbreaks of known plant pests or diseases, expansion of their geographical and/or host range, etc.), with proper multilingual definitions. Objective 3 is to develop and test a multilingual ontology for the global identification of emerging new plant health threats (e.g. emergence of new plant pests or diseases, new virulent genotypes with crop resistance breaking or pesticides resistance). Finally, Objective 4 is to analyse and test approaches and strategies for reporting the identified signals to the EFSA Units and experts through the MedISys interface, including mapping and geo-referencing.

To address Objective 1, the project provided an inventory of information sources relevant to monitor plant health threats through a systematic literature review methodology. Additionally, it has monitored information sources for the identification of new world-wide sources to determine how relevant the identified sources are and to help distinguish between adequate and inadequate websites. This task has been running continuously during the entire project. Potential sources have been evaluated to provide a quality assessment (usefulness) of identified information sources from the point of view of plant health threats detection.

The information sources selection process applied to achieve these goals is based on two methods: the direct and indirect methods. The direct method consists in identifying what is already known in the plant health threats domain, i.e. information sources recognised as relevant by the community, state of the art literature, etc. Usually, the most specific, efficient and accurate information sources in a search strategy are information collections and documents that are already known or that are recognised as relevant for a particular topic, like plant health. This kind of information sources is dealt by the direct method.

However, early warning information of plant health threats might be found first in nonofficial information sources, like general news or blogs. Therefore, it is necessary to include a second method capable of identifying this kind of information sources, which is called the indirect method. It is based on automatic Web searches using plant health threat keywords. With this method it is possible to identify information sources previously unknown for the plant health community at large but that can be also relevant, especially for new and re-emerging plant health threats.

Both methods were carried out in parallel. The final objective is to collect a list of relevant information sources in the plant health threats area that can be monitored by MedISys to detect both re-emerging and new emerging plant health threats.

The previous Web searches, and in general all the knowledge about plant health threats captured during the project, is organised using an ontology. An ontology is a formal representation of domain knowledge that can be easily computerised. The Plant Health Threat Ontology is a conceptual model to structure data about plant pests and diseases. The conceptual model provides the concepts and relations to be used to describe pests and their relations among them and to other related concepts like symptoms, hosts, vectors, crops, etc. The dataset contains the descriptions for the selected pests based on the previous conceptual model.

However, the main feature provided by the ontology in relation with known threats is the ability to organise for each one all the labels used to name in different languages, specifically the 10 languages the project focuses on: English, Spanish, Italian, French, German, Dutch, Portuguese, Chinese, Russian and Arabic. This is the information used to generate the MedISys categories for known threats, which addresses Objective 2. The core elements of the ontology and their relationships from a

conceptual point of view have been implemented using Semantic Web technologies, concretely the Web Ontology Language (OWL).

In addition to known threats, the project also aimed to monitor unknown plant health threats, as requested in Objective 3. Consequently, the project also explores ways to generate MedISys categories that do not use threat names as a selection criterion. Two approaches have been followed in this regard.

First of all, a more classical approach based on a manually curated category containing generic terms related to plant health threats, plus some negative terms making it possible to avoid irrelevant news items, e.g. those related to human health drugs.

Second, an alternative approach based on generating categories that include terms related to the threat but not its name. The ontology models symptoms, plant parts and vectors for 7 of the most active threats and they have been used to generate 7 categories just based on symptom expressions.

Finally, to address Objective 4, the MedISys website and alerting service based on e-mails were analysed. The MedISys website corresponds to the existing web pages that constitute the system interface for end-users, so that users can browse the different categories defined in MedISys and the news items captured by them. Features related to georeferencing, statistics per country and over time, or to export maps, were studied when interacting with the MedISys categories generated by the project. Following common practice in the user experience community, 14 user tasks were defined. They are based on what is available from the MedISys user interface for end-users and the experience gained by plant health experts while using it. These tasks characterise typical information needs that a user would like to satisfy using the features provided by the project.

On the other hand, the alerting service based on e-mails is a MedISys subscription based service that facilitates tracking plant health news. The user can register to receive periodic e-mails about news items captured by the categories he is interested in. A user study has been also conducted to evaluate the usefulness of this alerting service. A set of 29 plant health experts were subscribed to this service for the first 47 categories generated during the project. At the end of the project, 14 of them filled a survey, included in this report, about their impressions about the service.

Overall, from these evaluations, it can be inferred that the news items captured by the MedIsys categories defined during the project are relevant for plant health experts. The only caveat that was observed was with users receiving the daily e-mail alert. Most of them found it too frequent and including too many items. This problem can be now easily addressed by users themselves, who can define e-mail frequency and register just for the plant health threats they are interested in.

Table of contents

1. Introduction

1.1. Background and Terms of Reference as provided by the requestor

This contract was awarded by EFSA to:

Contractor: Universitat de Lleida (UdL), as coordinator, and Institute for Food and Agricultural Research and Technology (IRTA), subcontracted.

Contract title: Development and testing of the media monitoring tool MedISys for the monitoring, early identification and reporting of existing and emerging plant health threats

Contract number: OC/EFSA/PLH/2013/02

1.2. Interpretation of the Terms of Reference

The four specific objectives of the contract resulting from the procurement procedure are as follows:

- Objective 1: collate new and appropriate media and information sources (e.g. journals, magazines, webpages, etc.) at global level, to be appended to MedISys for the screening of plant health threats,
- Objective 2[1]: develop and test approaches and strategies to monitor re-emerging plant health threats on global and regional scales (e.g. new outbreaks of known plant pests or diseases, expansion of their geographical and/or host range, etc.), with proper multilingual definitions, to be appended to MedISys to the list of plant health alerts,
- Objective 3: develop and test a multilingual ontology for the global identification of emerging new plant health threats (e.g. emergence of new plant pests or diseases, new virulent genotypes with crop resistance breaking or pesticides resistance), to be appended to MedISys to the list of plant health alerts,
- Objective 4: analyse and test approaches and strategies for reporting the identified signals to the EFSA Units and experts through the MedISys interface, including mapping and geo-referencing (i.e. reports or alerts targeted to the EFSA Scientific Panel on Plant Health, the EFSA Scientific Network for plant health risk assessment and the EFSA Standing Working Group on emerging risks).

The contractor was trained at the onset of and during the project on the use of the MedISys tool and was provided access to MedISys in the execution of the tasks related to the contract resulting from this procurement procedure. Once trained, the contractor was capable of working autonomously in MedISys.

During the project implementation, the contractor worked in close liaison with EFSA and the JRC, the latter provided the training, platform access and maintenance and the technical support for the tasks related to the implementation of MedISys in the plant health area.

For a smooth progress of the project, a steering committee, composed by EFSA, the JRC and the contractor, monitored the project, with periodical meetings.

Deliverables: In total four Interim Reports (IR1-4) and three MS Excel spreadsheet (ES1-4), along with a consolidated Final Report (FR) addressing the four specific objectives and summarising the information described in the Interim Reports and the MS Excel spreadsheets, were prepared by the contractor.

[1] Objectives 2 and 3 have been interchanged as requested and agreed in the "Agreement to the request for amendment No. 1 to OC/EFSA/PLH/2013/02-CT1" dated May 20th, 2014 and signed by Marta Hugas, acting head of RASA Department.

Timelines: The above-mentioned deliverables were submitted as planned to EFSA within 30 months after the start of the project (i.e. from the date of the signature of the contract as a result of this procurement procedure):

A final report was delivered within 30 months.

2. Data and Methodologies

2.1. Data Model Conceptualisation

The Plant Health Threat Ontology is a conceptual model to structure data about plant pests and diseases. The conceptual model provides the concepts and relations to be used to describe pests and their relations, among them and to other related concepts like symptoms, hosts, vectors, crops, etc. The dataset contains the descriptions for the selected pests based on the previous conceptual model.

The next subsections describe all the steps followed to generate the ontology, starting from the list of agreed pests, the integration of this list with a reference dataset of taxons (currently the UniProt Taxonomy dataset) and the enrichment of this basic ontology with different pest names and related taxons. These concepts and data are, when possible, reused from existing datasets and ontologies, as described below. Finally, a list of all the datasets and ontologies used is included at the end of this section.

2.1.1. Ontology Design

The core of the ontology is the "Pest or Disease" concept that represents a plant health threat and is then linked to all the relevant concepts that help capturing relevant information about the threat, as shown in Figure 1. The threat is linked to crops or hosts and vectors. All these entities will be kinds of Taxon, the way taxa is defined in the UniProt Taxonomy dataset, and correspond to a species as detailed below. For instance, a plant, an insect, a virus, etc. Threats are also linked to Symptoms Expressions, which mainly connect symptoms to the plant parts that they affect, as also detailed later.

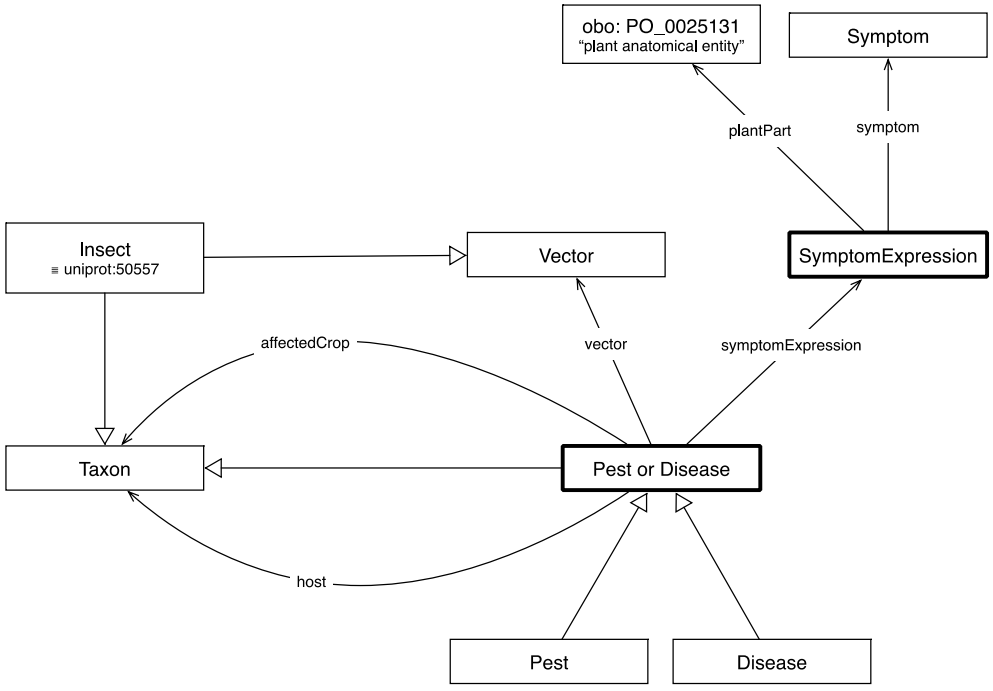


Figure 1: Core Plant Plant Health Threat Ontology design

These are the main datasets used to help populating the Plant Health Threat Ontology:

Taxon

Instances of the concept "Taxon" are reused from the UniProt Taxonomy dataset, a database that is maintained by the UniProt group and is based on the NCBI taxonomy database. Organisms are classified in a hierarchical tree structure. The UniProt Taxonomy database contains every node (taxon) of the tree. Instances of Taxon are related to the "Pest or Disease" concept through the properties "crop" and "host". The ontology is available from: http://www.uniprot.org/taxonomy/

Plan Part

Instances of the concept "plant anatomical entity" are reused from the Plant Ontology (PO). This ontology describes plant anatomy, morphology and stages of development for all plants. The goal of the PO is to establish a semantic framework for meaningful cross-species queries. The ontology is available from: http://www.cropontology.org/ontology/PO/Plant%20Ontology

The previous conceptualisation of the ontology has been implemented using Semantic Web technologies, concretely the Web Ontology Language (OWL). The ontology is shown in Figure 2.

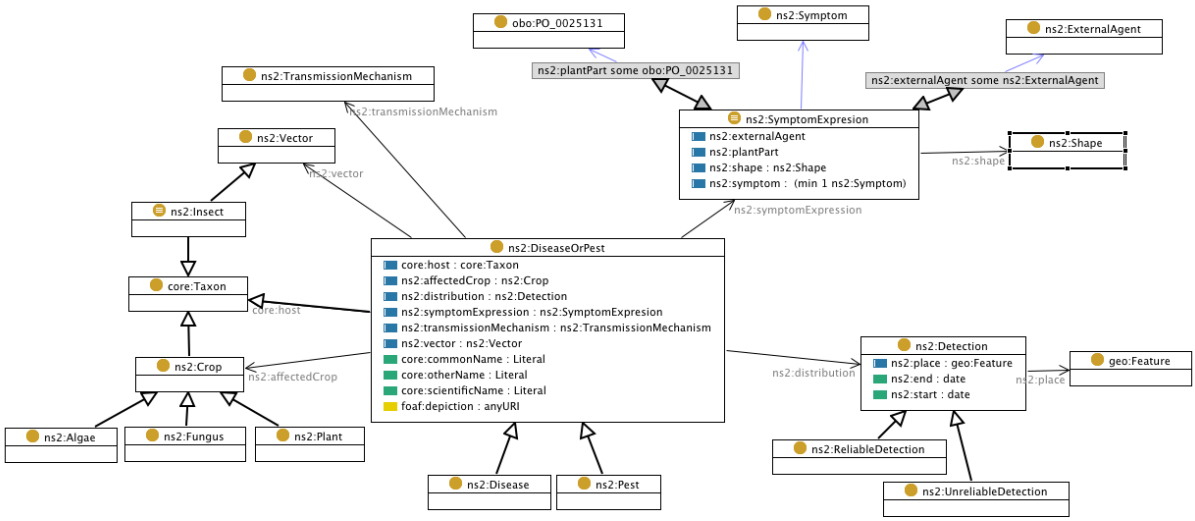


Figure 2: Core Plant Plant Health Threat Ontology implementation

2.1.2. Pests and Diseases Modelled using the Ontology

The initial set of pests under consideration to generate the categories included the 47 pests listed in Table 1. The table lists their scientific name and the corresponding category, which was used to organise the alert depending on the kind of organism and based on the broad concepts *bacteria*, *fungi*, *insecta*, *mollusca*, *nematoda*, *oomycetes*, *viroid* and *virus*. These categories are used to organise the resulting categories in the MedISys website.

Table 1: Plant pests and diseases for which MedISys categories based on threat names were generated. There is also a category for *Trichilogaster acaciaelongifoliae*, a gall wasp used as a biological control agent of an invasive alien plant

Category	Scientific Name
bacteria	Candidatus liberibacter
bacteria	Xylella fastidiosa
fungi	Ceratocystis fagacearum
fungi	Diplocarpon mali
fungi	Geosmithia morbida
fungi	Heterobasidion irregulare
fungi	Hymenoscyphus fraxineus
fungi	Monilinia fructicola
fungi	Thecaphora solani
fungi	Tilletia indica
insecta	Agrilus coxalis auroguttatus
insecta	Agrilus planipennis
insecta	Anastrepha ludens
insecta	Anomala orientalis
insecta	Anoplophora glabripennis
insecta	Bactrocera tryoni
insecta	Diabrotica virgifera
insecta	Rhagoletis cingulata
insecta	Rhagoletis fausta
insecta	Rhagoletis indifferens
insecta	Rhagoletis mendax
insecta	Rhagoletis ribicola
insecta	Rhagoletis suavis
insecta	Rhynchophorus ferrugineus
insecta	Spodoptera eridania
insecta	Spodoptera frugiperda
insecta	Spodoptera litura
insecta	Tecia solanivora
insecta	Thrips palmi
insecta	Trichilogaster acaciaelongifoliae
mollusca	Pomacea
nematoda	Bursaphelenchus xylophilus
nematoda	Nacobbus aberrans
nematoda	Punctodera chalcoensis
oomycetes	Phytophthora ramorum

viroid	Potato spindle tuber viroid
virus	Andean potato latent virus
virus	Andean potato mottle virus
virus	Cowpea mild mottle virus
virus	Euphorbia mosaic virus
virus	Lettuce infectious yellows virus
virus	Peach rosette mosaic virus
virus	Pepper mild tigre virus
virus	Potato black ringspot virus
virus	Potato virus T
virus	Strawberry vein banding virus
virus	Tobacco ringspot virus

At the end of the project, July 2016, it has been possible to extend the set of names pests and diseases with the 70 additional ones listed in Table 2. With this addition, now 117 of the original candidate list of 140 pests have been processed. They correspond to all the pests that have an equivalent in UniProt Taxon and for which it has been possible to retrieve scientific and common names in different languages. Those that for which we have not found a mapping are listed in Table 3.

The 140 pest under consideration, as suggested in the Call for Tender and agreed during project meetings with EFSA, are:

- EPPO Alert (as January 2014)
- 2000/29 1-A-1 (as January 2014)
- EU Emergency Measures (as January 2014)
- EFSA's additional suggestions: *Xylella fastidiosa*, *Hymenoscyphus fraxineus* (Ash Dieback Disease), *Agrilus planipennis* (Emerald Ash Borer), *Anoplophora glabripennis*, *Candidatus Phytoplasma pruni*, *Trichilogaster acaciaelongifoliae* (a weed biocontrol agent) and *Candidatus liberibacter* (Citrus Greening).

Table 2: Additional pests and diseases added by the end of the project

Category	Scientific Name
bacteria	Acidovorax citrulli
bacteria	Candidatus Arsenophonus phytopathogenicus
bacteria	Pseudomonas syringae pv. actinidiae
bacteria	Pseudomonas syringae pv. aesculi
fungi	Botryosphaeria laricina
fungi	Chrysomyxa arctostaphyli
fungi	Gibberella circinata
fungi	Mycosphaerella laricis-leptolepidis
fungi	Mycosphaerella populorum

fungi	Phellinus weirii
fungi	Phymatotrichopsis omnivora
fungi	Septoria malagutii
fungi	Sirococcus tsugae
fungi	Stagonosporopsis andigena
insecta	Anastrepha fraterculus
insecta	Anastrepha obliqua
insecta	Anastrepha suspensa
insecta	Aproceros leucopoda
insecta	Aromia bungii
insecta	Arrhenodes minutus
insecta	Bactrocera cucurbitae
insecta	Bactrocera dorsalis
insecta	Bactrocera tsuneonis
insecta	Bactrocera zonata
insecta	Ceratitis quinaria
insecta	Ceratitis rosa
insecta	Conotrachelus nenuphar
insecta	Dacus ciliatus
insecta	Diabrotica barberi
insecta	Diabrotica undecimpunctata howardi
insecta	Diabrotica undecimpunctata undecimpunctata
insecta	Dryocosmus kuriphilus
insecta	Epitrix
insecta	Euphranta canadensis
insecta	Haplaxius crudus
insecta	Helicoverpa zea
insecta	Liriomyza sativae
insecta	Myiopardalis pardalina
insecta	Nemorimyza maculosa
insecta	Neoceratitis cyanescens
insecta	Pityophthorus juglandis
insecta	Polygraphus proximus
insecta	Pseudopityophthorus minutissimus
insecta	Pseudopityophthorus pruinosus
insecta	Rhagoletis completa
insecta	Rhagoletis pomonella
insecta	Thaumastocoris peregrinus
insecta	Xylosandrus crassiusculus
nematoda	Heterodera zeae

nematoda	Longidorus diadecturus
nematoda	Meloidogyne ethiopica
nematoda	Xiphinema californicum
phytoplasma	Candidatus Phytoplasma fragariae
phytoplasma	Candidatus Phytoplasma pruni
phytoplasma	Candidatus Phytoplasma solani
phytoplasma	Elm yellows phytoplasma
phytoplasma	Peach rosette phytoplasma
phytoplasma	Peach X-disease phytoplasma
phytoplasma	Peach yellows phytoplasma
viroid	Tomato apical stunt viroid
virus	American plum line pattern virus
virus	Arracacha virus B
virus	Bean golden mosaic virus
virus	Blueberry leaf mottle virus
virus	Cherry rasp leaf virus
virus	Hosta virus X
virus	Peach mosaic virus
virus	Pepino mosaic virus
virus	Squash leaf curl virus
virus	Tomato mottle virus

Table 3: The 23 pests and diseases without direct mapping to UniProt Taxon

```
Arracacha virus B (Oca Strain), Carneocephala fulgida, American cherry rasp leaf virus, Chrysophtharta bimaculata,
Draeculacephala minerva, Elm phloem necrosis mycoplasm, Florida tomato virus, Melampsora farlowii, Neoleucinodes elegantalis,
Ophiomyia kwansonis, American peach mosaic virus, Peach phony rickettsia, Peach rosette mycoplasm, Peach X-disease mycoplasm,
Peach yellows mycoplasm, Phyllosticta solitaria, American raspberry leaf curl virus, Rhacochlaena japonica, Scaphoideus
luteolus, Strawberry latent 'C' virus, Strawberry witches' broom mycoplasm, Tomato apical stunt pospiviroid, Trechispora
brinkmannii
```

2.1.3. Symptoms Expressions

The symptom expression part of the ontology focuses on modelling the symptoms associated to plant pests and diseases, and how they are usually expressed, for instance affecting a specific plant part. As previously mentioned, the plant parts were reused from the Plant Ontology. On the other hand, the range of symptoms that are considered in the ontology were reused from CABI, which provides a form to capture the symptoms associated to a plant health threat with a predefined set of symptoms and plant parts they affect.

The symptoms modelled by the ontology, together with their available translations for the selected languages, can be retrieved using the following SPARQL query:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX pht: <http://rhizomik.net/ontologies/PlantHealthThreats#>
SELECT ?sym
  (GROUP_CONCAT(DISTINCT ?en; separator = ", ") AS ?all_en)
  (GROUP_CONCAT(DISTINCT ?es; separator = ", ") AS ?all_es)
  (GROUP_CONCAT(DISTINCT ?it; separator = ", ") AS ?all_it)
  (GROUP_CONCAT(DISTINCT ?fr; separator = ", ") AS ?all_fr)
  (GROUP_CONCAT(DISTINCT ?de; separator = ", ") AS ?all_de)
  (GROUP_CONCAT(DISTINCT ?nl; separator = ", ") AS ?all_nl)
  (GROUP_CONCAT(DISTINCT ?pt; separator = ", ") AS ?all_pt)
  (GROUP_CONCAT(DISTINCT ?ru; separator = ", ") AS ?all_ru)
  (GROUP_CONCAT(DISTINCT ?ar; separator = ", ") AS ?all_ar)
  (GROUP_CONCAT(DISTINCT ?zh; separator = ", ") AS ?all_zh)
WHERE {
  ?sym a pht:Symptom
  OPTIONAL { ?sym rdfs:label ?en FILTER(lang(?en)='en') }
  OPTIONAL { ?sym rdfs:label ?es FILTER(lang(?es)='es') }
  OPTIONAL { ?sym rdfs:label ?it FILTER(lang(?it)='it') }
  OPTIONAL { ?sym rdfs:label ?fr FILTER(lang(?fr)='fr') }
  OPTIONAL { ?sym rdfs:label ?de FILTER(lang(?de)='de') }
  OPTIONAL { ?sym rdfs:label ?nl FILTER(lang(?nl)='nl') }
  OPTIONAL { ?sym rdfs:label ?pt FILTER(lang(?pt)='pt') }
  OPTIONAL { ?sym rdfs:label ?ru FILTER(lang(?ru)='ru') }
  OPTIONAL { ?sym rdfs:label ?ar FILTER(lang(?ar)='ar') }
  OPTIONAL { ?sym rdfs:label ?zh FILTER(lang(?zh)='zh') }
} GROUP BY ?sym ORDER BY ?sym
```

The result is the following table, which features 37 different symptoms, with the associated terms in English and other languages when available:

English	Spanish	Italian	French	German	Dutch	Portuguese	Russian	Arabic	Chinese
abnormal fall, premature fall	caída prematura	caduta prematura	chute anormale, chute prématurée	verfrüht herbst	voortijdige val	queda anormal, queda prematura			早期脱落
abnormal patterns, chlorotic rings	anillos cloróticos, listado, líneas cloróticas, mosaico	anelli clorotiche	anneaux chlorotiques	chlorotisch Ringe	chlorotische ringen	padrões anormais	хлоротичные кольца		褪绿环
abnormal shape, malformation, distortion	forma anormal, malformación	malformazione	déformation, forme anormale, malformation	missbildung	misvorming	distorção, malformacão		تشوه	畸形
boring, drilling, internal feeding, mining, tunneling		perforazione, foratura	forage	bohren	boring, boren			حفر	钻孔
canker	antracnosis, cancro, chancro	antracnosi, cancrena	chancre	baumkrebs		antracnose	антракноз		溃疡
chlorosis	clorosis	clorosi	chlorose	chlorose	bleekzucht	cloroses	хлороз	الاخضلوروز دم في قر	萎黄
colour inversion,	color	inversione	inversion	farbinvertie	kleurinve	inversão	инверсия	انعكاس	颜色

color inversion	invertido	dei colori	couleur	rung	rsie	da cor	цвета	ال لون	反转
curling, curl	enrollamiento	arricciatura	enroulement	eisschießen			вьющийся	م تجعد	冰壶
dieback	muerte regresiva, muerte descendente	deperimento	dépérissement		afstervin g		отмирание	ال سـ قم	顶枯病
discoloration, discolouration	decoloración		decoloration	verfärbung	verkleuri ng	descoloraç ão	обесцвечивание	ت غ ي ير ال لون	变色
dwarfing	enanismo	nanismo	nanisme	zwergwuchs	dwerggroe i		карликовость	ت قزم	矮化
early senescence, premature senescence		senescenza precoce	sénescence prématurée	vorzeitige vergreisung	voortijdige verouderi ng	senescência antecipada	преждевременное старение	ال شـ يخ وخة ال م بكر ة	早衰
empty	vacío	vuoto	vide						
feeding									
frass				frass	frass		экскременты		虫粪
gummosis	gomosis		gomose				гоммоза		流胶病
lesion, lesions	lesiones, lesión	lesioni	blessures	läsionen	laesies		поражения	الأفـ ات	病变
mottled, mottle	motear, moteado	macchia	marbrure	klecks	vlek		пятнистость		斑点
mummification, wrinkled, hard skin	piel dura, momificación, arrugada, arrugado	essiccazione, pelle dura, mummificazione, rugoso	séchage, peau dure, momification, ridée	getrocknet, trocknung, verhärtete Haut, mumifizierung, zerknittert	gedroogd, drogen, harde huid, mummificatie, gerimpeld	de secagem, pele dura, mumificaçã o, enrugado	сушеные, сушка, жесткие кожи, мумификация, сморщенный	ال مجـ فف ة, ت جـ في ف, ال جـ لد ال ثابـ ت, ال متحـ ني ط, ت جعد	干, 干燥, 坚硬的皮肤, 皱
dead, death, necrosis	muerte, muerto, necrosis, muerta	necrosi, morte	nécrose, mort	nekrose, tod	necrose, dood		некроз, мертвых, смерть	,نـ خر موت	壞死, 坏疽, 坏死, 死亡
odour	olor	odore	odeur	geruch	geur	odor	запах	رائـ حة	气味
premature drop	caida prematura		chute prématurée						
premature ripening	maduración prematura		maduration précoce, maturation prématurée						

reddening	enrojecimiento	arrossamenti	rougissement	rötung	roodheid	vermelhidão	покраснение	احمرار	泛红
reduced size, smaller	tamaño reducido, más pequeño								
resinosis			résinose						
roll, rolling	enrollado		enroulement, roulement						
rosetting	roseta		rosette	rosetten					
rot, rotting	putrefacción, podredumbre, pudrición	marciume	pourriture	fäulnis, fäule	rot	podridão	гниль	ع فن	腐烂
burn, scorch	quemadura	bruciatura	brûlure, brûlée	versengen	verschroeien	queimadura	сожженный	حرق	烧伤
splitting	agrietado, agrietada		scission						
stunting	enanismo, retraso en el crecimiento		nanisme, rabougrissement						
thicker	más grueso		plus épais						
fallen, toppled, falling	caído, derribado	caduto	tombant, tombé			derrubado, queda	падший	ساق ط	堕落
rooted out, uprooted	desarraigado	sradicato	déraciné	entwurzelt	ontworteld	arrancado, desenrraizado	выкорчевали		连根拔起
wilt, wilting	marchitamiento	avvizzimento	flétrissement	welk	verwelking	murchamento	увядание	ذبـ ول	萎蔫
yellowing	amarillamiento	ingiallimento	jaunissement	vergil	vergeling	amarelecimento	желтеет, пожелтение	ا صـ فرار	片黄化

Moreover, symptoms were associated to plant parts. The ontology has a selection of the relevant plant parts associated to symptoms expressions as organised in the CABI form. Though additional, more specific, plant parts can be also used in symptom expressions, this reduced set of the more relevant ones from the CABI perspective was included in the ontology together with their translations to the selected languages.

The following SPARQL query can be used to retrieve them:

```
PREFIX pht: <http://rhizomik.net/ontologies/PlantHealthThreats#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
SELECT ?plant_part
  (GROUP_CONCAT(DISTINCT ?en; separator = ", ") AS ?all_en)
```

```
  (GROUP_CONCAT(DISTINCT ?es; separator = ", ") AS ?all_es)
  (GROUP_CONCAT(DISTINCT ?it; separator = ", ") AS ?all_it)
  (GROUP_CONCAT(DISTINCT ?fr; separator = ", ") AS ?all_fr)
  (GROUP_CONCAT(DISTINCT ?de; separator = ", ") AS ?all_de)
  (GROUP_CONCAT(DISTINCT ?nl; separator = ", ") AS ?all_nl)
  (GROUP_CONCAT(DISTINCT ?ru; separator = ", ") AS ?all_ru)
  (GROUP_CONCAT(DISTINCT ?ar; separator = ", ") AS ?all_ar)
  (GROUP_CONCAT(DISTINCT ?zh; separator = ", ") AS ?all_zh)
WHERE {
  ?plant_part a obo:PO_0025131
  OPTIONAL { ?plant_part ?p ?en FILTER(lang(?en)='en') }
  OPTIONAL { ?plant_part ?p ?es FILTER(lang(?es)='es') }
  OPTIONAL { ?plant_part ?p ?it FILTER(lang(?it)='it') }
  OPTIONAL { ?plant_part ?p ?fr FILTER(lang(?fr)='fr') }
  OPTIONAL { ?plant_part ?p ?de FILTER(lang(?de)='de') }
  OPTIONAL { ?plant_part ?p ?nl FILTER(lang(?nl)='nl') }
  OPTIONAL { ?plant_part ?p ?pt FILTER(lang(?pt)='pt') }
  OPTIONAL { ?plant_part ?p ?ru FILTER(lang(?ru)='ru') }
  OPTIONAL { ?plant_part ?p ?ar FILTER(lang(?ar)='ar') }
  OPTIONAL { ?plant_part ?p ?zh FILTER(lang(?zh)='zh') }
} GROUP BY ?plant_part
```

The result is shown in the following table, which shows these 6 generic plant parts associated to symptom expressions and their translations to the selected languages:

English	Spanish	Italian	French	German	Dutch	Russian	Arabic	Chinese
fruit	fruta, frutas, fruto	frutta, frutto	fruits	frucht, obst	vrucht	плод, фрукты	ثمرة, الفاكهة	果实, 水果
plant, tree, whole plant	árbol, planta entera, toda la planta	pianta	arbre, plante	pflanzen	boom, plant	Растения	نبات	植物
bud, sprout	brote, yema	germoglio	bourgeon	knospe, sprießen	kiem	бутон	برعم	新芽, 芽
stem	tallo, vástago	fusto	tige		stam	стебля	الجذعية, والساق	茎
seed, seeds	semilla, semillas	seme, semi	graine	same	zaad	Семя	بذرة	種子
leaf, leaves	hojas, hoja	foglia, foglie	feuille, feuilles	blatt, blätter	blad, bladeren	лист, листья	ورقة, أوراق, ورقة نبات	叶, 葉

2.1.4. Linking Selected Pests to a Reference Dataset

In order to generate an ontology, the list of pest names resulting from integrating the reference pest lists was aligned with a reference dataset that provides unambiguous identifiers for each pest. This process is called *Reconciliation*. Different datasets available as open data were evaluated and used during this reconciliation process.

The selected dataset to provide the identifiers for pests was the UniProt Taxonomy dataset. These are the details about the dataset:

- Name: UniProt Taxonomy
- Availability (as part of the UniProt RDF Distribution)
    - ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/rdf/README
- Download
    - http://www.uniprot.org/taxonomy/?query=&force=yes&format=rdf
    - ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/rdf/taxonomy.rdf.gz
- SPARQL EndPoint: http://omediadis.udl.cat:8890/sparql
    - Where the dataset has been loaded so it is available for querying using the SPARQL query language for semantic data.
- Graph: http://purl.uniprot.org/taxonomy/
    - Identifier for the data graph from where UniProt Taxonomy data can be retrieved from the previous SPARQL EndPoint.
- Reconciliation properties:
    - http://purl.uniprot.org/core/scientificName
    - http://purl.uniprot.org/core/otherName
    - http://purl.uniprot.org/core/commonName
        - These three properties provide taxon names that can be matched with the pest names in order to retrieve the associated taxon identifier. This identifier will be used as the pest identifier after reconciliation.
- Other relevant properties
    - http://purl.uniprot.org/core/host
        - This property, though not frequent in the UniProt dataset, provides for some taxons a link to their host. This can be reused to fill the "host" property of pests in the ontology.

Reconciliation Process

The UniProt Taxonomy dataset was used as the source of identifiers for the selected pests. To automatize the process of checking for each pest name the matching taxon in UniProt, the LODRefine tool was used. This tool was used to match the labels associated to a taxon against the pest names in the selected pest lists. In order to do so, the taxon names defined by in UniProt dataset (*scientific name*, *common name* and *other name*) were copied to labels so they could be all used during the reconciliation process. To do that the following SPARQL Update command was used:

```
PREFIX core <http://purl.uniprot.org/core/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

WITH <http://purl.uniprot.org/taxonomy/>
INSERT { ?r rdfs:label ?label }
WHERE {
  ?r a core:Taxon; ?p ?label
  FILTER(?p=core:commonName||?p=core:scientificName||?p=core:otherName) }
```

The reconciliation process was then performed automatically by LODRefine, though it was also possible to interactively redefine the matches automatically detected. From these matches, it was then possible to associate to each pest a taxon identifier from UniProt Taxonomy, which looks like *http://purl.uniprot.org/taxonomy/414338*, as shown in the LODRefine screenshot in Figure 3.

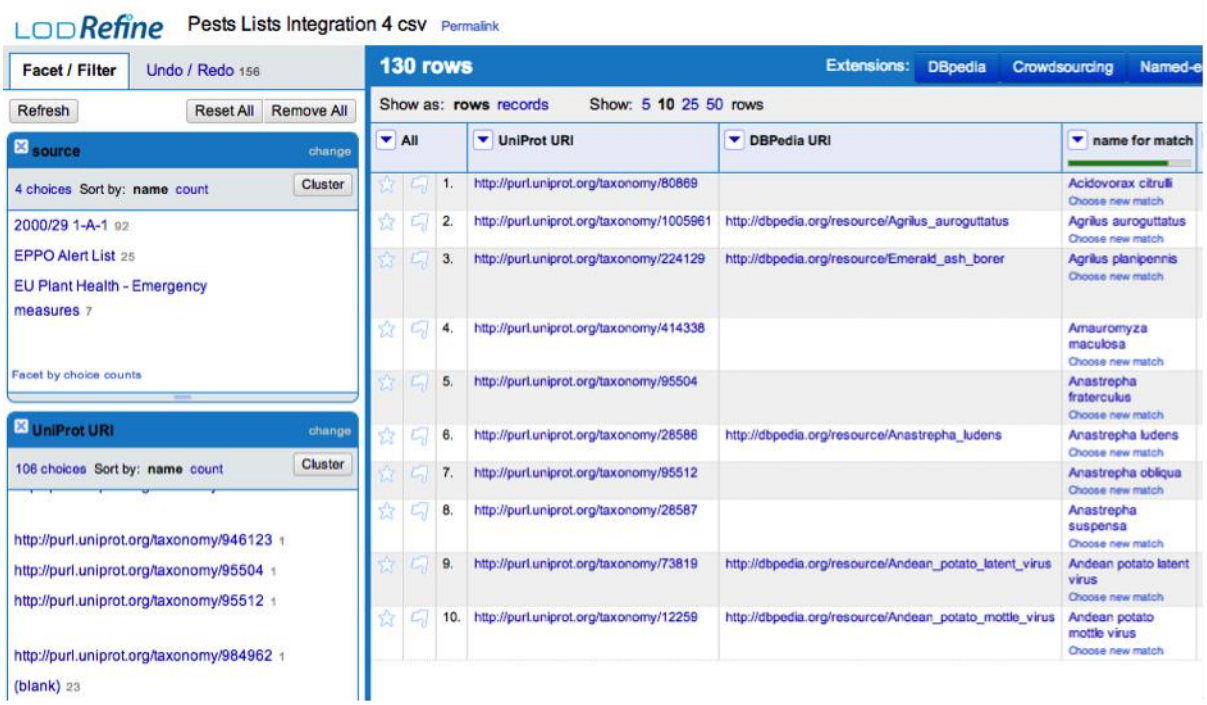


Figure 3: Screenshot of the LODRefine reconciliation tool after associating pests to taxon identifiers (UniProt URIs) and also DBPedia identifiers (DBPedia URIs)

In addition to reconciling the pest list against UniProt, LODRefine was also used to try to reconcile them against DBPedia, a semantic dataset generated from Wikipedia. When it was possible to link the pest to an entry in the Wikipedia (through DBPedia), the pest was linked to the corresponding data making it possible to retrieve from Wikipedia information about alternative pest names, including other languages, and other related entities. Some examples of DBPedia identifiers (URIs), for those pests that were matched with DBPedia, are also shown in Figure 3.

Generating the Semantic Representation of Selected Pest

Finally, after completing the reconciliation process against UniProt Taxonomy and DBPedia, LODRefine was used to generate the semantic data that constituted the starting point to build the Plant Health Threat Ontology.

LODRefine provides a template building service that facilitates generating RDF semantic data from tabular data (e.g. spreadsheets), in our case the pests list. Figure 4 shows the template used to generate the semantic data for the Plant Health Threat Ontology.

As it is shown, the UniProt URI was used as the main identifier for the pests. It was then connected to the DBPedia URI, if it was present, as an equivalent identifier. The values of the corresponding *pest name* and *other name* cells were linked to each pest as labels. The output semantic data also kept track of the source list from where the pest was collected using the *source* property. Finally, an

additional *comment* property was included to associate the pest to comments if they were present in the original tabular data.

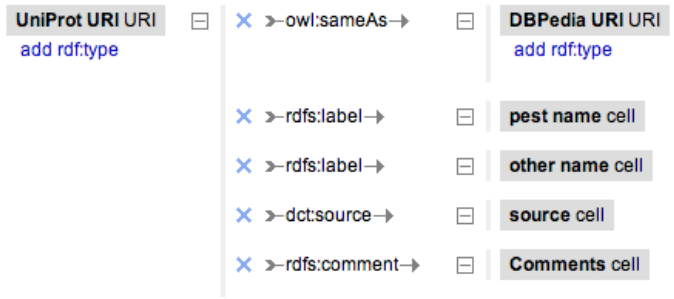


Figure 4: LODRefine template defined to generated semantic data from the input pests list available as tabular data (spreadsheet)

Table 4 shows an example of the resulting semantic data, based on the Resource Description Framework (RDF) standard.

Table 4: Example of semantic data for the *Agrilus planipennis* pest from EFSA proposals list

```
<http://purl.uniprot.org/taxonomy/224129> a pests:Pest;
       rdfs:label "Agrilus planipennis", "Emerald ash borer";
       owl:sameAs "http://dbpedia.org/resource/Emerald_ash_borer";
       dct:source "EFSA";
       rdfs:comment "Check spread from Russia towards EU".
```

Generating Ontology Hierarchical Structure from Taxonomy

The semantic data obtained so far, based on the previous template, constituted the core of the Plant Health Threat Ontology dataset. It was loaded into a semantic data repository, in our case a Virtuoso Open Source deployment. The semantic data generated as a result of the process detailed in the previous section was loaded using the following commands:

```
SPARQL CLEAR GRAPH <http://rhizomik.net/pests/>

DB.DBA.TTLP_MT(file_open('/opt/virtuoso-db/pests/Pests-Lists-
Integration.ttl'),'','http://rhizomik.net/pests/', 255);
```

The dataset was then enriched with information from other semantic sources. The first enrichment step, detailed in this section, was to reuse the taxon hierarchy structure from the UniProt Taxonomy dataset. This dataset was also used to retrieve the scientific, common and alternative names defined in UniProt and information about hosts, if it was present.

The SPARQL Update query to perform the per-pest enrichment with names and information about host was:

```
PREFIX taxon: <http://purl.uniprot.org/core/>
PREFIX pests: <http://rhizomik.net/ontologies/pests.owl#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
```

```
INSERT {GRAPH <http://rhizomik.net/pests/> {
               ?r a owl:Class ;
                      taxon:scientificName ?scientific; taxon:otherName ?other;
                      taxon:commonName ?common; taxon:host ?host } }
WHERE {
       GRAPH <http://rhizomik.net/pests/> {
               ?r a pests:Pest }
       GRAPH <http://purl.uniprot.org/taxonomy/> {
               ?r a taxon:Taxon
               OPTIONAL { ?r taxon:scientificName ?scientific }
               OPTIONAL { ?r taxon:otherName ?other }
               OPTIONAL { ?r taxon:commonName ?common }
               OPTIONAL { ?r taxon:host ?host } }
}
```

Then, the second enrichment step was to populate the taxonomy hierarchy from identified pests up through the taxon hierarchy in UniProt taxonomy. This way, the ontology was filled with all the ancestor taxons starting from the identified pests.

This was also done querying the Virtuoso semantic datastore using the following SPARQL Update query, which copied the ancestor taxons from UniProt to the Plant Health Threat Ontology together with their names and hosts:

```
PREFIX taxon: <http://purl.uniprot.org/core/>
PREFIX pests: <http://rhizomik.net/ontologies/pests.owl#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

INSERT {GRAPH <http://rhizomik.net/pests/> {
       ?ancestor a owl:Class; rdfs:subClassOf ?parent;
               rdfs:label ?scientific;
               taxon:scientificName ?scientific; taxon:otherName ?other;
               taxon:commonName ?common; taxon:host ?host }
WHERE {
       GRAPH <http://rhizomik.net/pests/> { ?r a pests:Pest }
       GRAPH <http://purl.uniprot.org/taxonomy/> {
               ?r rdfs:subClassOf+ ?ancestor.
               ?ancestor a taxon:Taxon.
               OPTIONAL { ?ancestor rdfs:subClassOf ?parent }
               OPTIONAL { ?ancestor taxon:scientificName ?scientific }
               OPTIONAL { ?ancestor taxon:otherName ?other }
               OPTIONAL { ?ancestor taxon:commonName ?common }
               OPTIONAL { ?ancestor taxon:host ?host }
}
```

With the additional information gathered from UniProt, it was possible to draw the taxon hierarchy for all the collected pests. First of all, the semantic data for the hierarchical structure plus the labels for the nodes to be drawn was extracted using the following SPARQL query:

```
PREFIX taxon: <http://purl.uniprot.org/core/>
PREFIX pests: <http://rhizomik.net/ontologies/pests#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

CONSTRUCT {
```

```
        ?r a owl:Class; rdfs:subClassOf ?super; rdfs:label ?scientific.
        ?super a owl:Class; rdfs:label ?scientific2.
}
FROM <http://rhizomik.net/pests/>
WHERE {
        ?r a owl:Class; rdfs:subClassOf ?super.
        OPTIONAL { ?r taxon:scientificName ?scientific }
        OPTIONAL { ?super taxon:scientificName ?scientific2 }
}
```

The resulting RDF semantic data was copied and pasted in the RDF2SVG[2] web service to generate a graphical representation of all the pests and their ancestor taxons as a hierarchy, as shown in Appendix D.

2.2. Identification of Media Sources for Monitoring

The information sources selection process applied to achieve the project goals was based on two methods: the direct and indirect methods. They were based on a standardised approach (Salaún and Flores, 2001) to identify and evaluate relevant information sources and to collect, report and analyse those that may be relevant to monitor emerging or re-emerging plant health threats.

The direct method, detailed in Section 2.2.1, consisted in identifying what was already established in the plant health domain, i.e. information sources recognised as relevant by the community, state of the art literature, etc. Usually, the most specific, efficient and accurate information sources in a search strategy are information collections and documents that are already known or that are recognised as relevant for a particular topic, like plant health. This kind of information sources was dealt by the direct method.

However, early warning information of plant health threats might be found first in nonofficial information sources, like general news or blogs, as already observed in previous EFSA reports (EFSA, 2012). Therefore, it was necessary to include a second method capable of identifying this kind of information sources, the indirect method detailed in Section 2.2.2. It was based on automatic Web searches using plant health threat keywords collected from the ontology presented in Section 2.1. With this approach, it was possible to identify information sources previously unknown for the plant health community at large but that were also relevant, especially for new and re-emerging plant health threats.

Both methods were carried out in parallel, as shown in Figure 5. The final objective was to collect a list of relevant information sources in the plant health threats area that can be monitored by MedISys to detect both re-emerging and new emerging plant health threats.

[2] RDF2SVG service, http://rhizomik.net/html/redefer/rdf2svg-form/

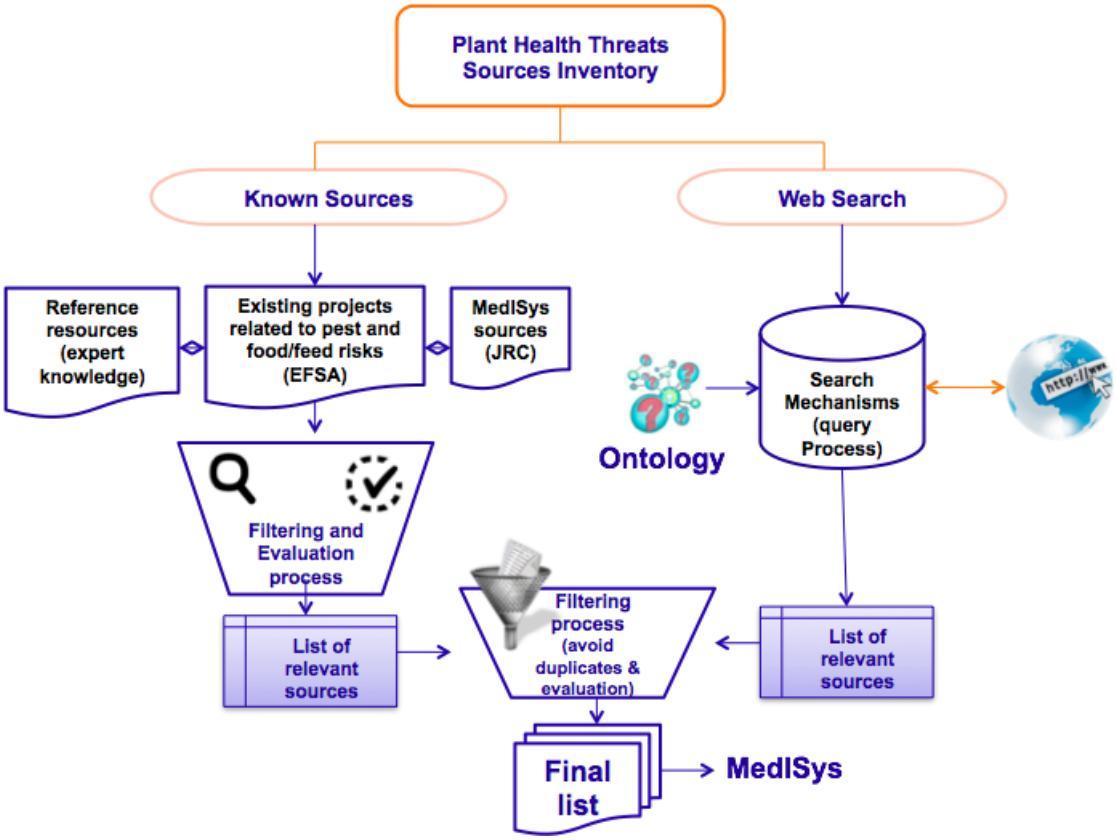


Figure 5: Overview of the sources selection methodologies. Direct method from known sources (left) and indirect method from Web search (right)

2.2.1. Direct Method: Manual Curation of Media Sources

This part of the methodology consists in the identification, review and evaluation of known sources to find relevant information sources related to plant health. Three types of sources collections were considered in this process:

- Collections of information sources well known among the plant health community. This includes sources already listed in the project proposal or those contributed during the project by IRTA members or organisations like EPPO. The proven relevance of these sources also served as a quality control for the sources obtained using the automatized indirect process.

- Repositories of information sources produced by previous projects related to pest and food or feed risks. This includes previous EFSA projects related to literature or information sources review. During the project, new projects providing such relevant collections were identified. These sources were also reviewed, like in the case of the PestLens project.

- The collection of sources already considered by MedISys as provided by JRC. This includes two kinds of sources: general media sources (with world coverage) and official sources or member states sources, though these sources focus mainly on the medical domain[3]. The main aim of considering existing sources from MedISys was to avoid selecting sources already monitored. Moreover, by reviewing these sources it was possible to find related sources that

[3] MedISys monitors human and animal infectious diseases, chemical, biological, radiological and nuclear (CBRN) threats, and recently food & feed hazards.

could be relevant for plant health, for example starting from a link to a website subsection related to human health it was possible find another subsection in the same site about plant health.

Review and Evaluation Process

During this process, each known source was reviewed and classified into one of four groups:

1. The first group corresponds to information sources not related to plant health. All sources in this group were immediately discarded.

2. The second group corresponds to sources related to plant health but whose information is static and mainly descriptive, for instance sources describing a known pest or those about laboratory experiments. These sources were also discarded because their lack of changes would make their monitoring very unproductive.

3. The third group are sources with relevant information about plant health but that cannot be monitored because they do not provide regular updates through a news section or feed (e.g. RSS). Some of these sources were finally selected, based on the information quality measures presented in this section, thought the absence of these features penalises them because it makes it harder for MedISys to monitor them. If these features, e.g. RSS, are not provided by the site, MedISys has to guess when the page has changed and it might happen that some updates are lost. This was observed, for instance, in the case of a BBC news item (http://www.bbc.com/news/science-environment-34120766) about *Xylella fastidiosa* inspired by an article by Steven Parnell, an EFSA PLH Panel Member. Though this was a relevant item and MedISys was monitoring the part of the BBC site the item appeared in, it was missed because there was no RSS.

4. Finally, the last group corresponds to news sources, alert mechanisms or news feeds focused on plant health threats, which are easily monitored by MedISys and thus were prioritised.

It is important to note that, for those information sources whose analysis identified several relevant subsections, each subsection was considered as an independent information source to be evaluated. These subsections have different URLs and content. For example, in an online newspaper with two different relevant sections, each of these sections is processed as a different information source. On the other hand, information sources that, though relevant, do not freely provide at least a summary of the news items, e.g. subscription-based sources, were not selected because they cannot be monitored by MedISys.

Information Quality Measures

Many different criteria have been proposed to measure the quality of online information sources about a specific topic or in a particular scientific domain[4] (Lee, et al. 2002; Naumann, 2002; Stvilia, 2007). For the selection of information sources about plant health for their monitoring using MedISys, the proposed process is based on an Information Quality Management (IQm) framework already used in previous EFSA projects such as Dataquest.

This framework takes into account two different dimensions of information sources: how sources are described using metadata (the Dublin core description detailed in Appendix A) and the analysis of their content (the data quality parameters detailed in Appendix B).

The review and evaluation process steps are:

[4] DACO Project, Dataquest Project, PRASSIS, etc.

- To review information sources to conclude whether or not they are related to plant health.
- To identify if these sources feature news sections, feeds or alert systems.
- To evaluate information sources metadata and content to measure their quality using the IQm framework.
- To select the information sources above a quality threshold defined by the information quality framework. These sources were included in deliverable ES1.

For the evaluation of information sources metadata, a complete description of each source including the following 14 metadata properties was taken into account:

1. **Title**: a name given to the source.
2. **Creator**: an entity primarily responsible for making the source.
3. **Subject**: the topic of the source.
4. **Description**: an account of the source.
5. **Publisher**: an entity responsible for making the source available.
6. **Contributor**: an entity responsible for making contributions to the source.
7. **Date**: a point or period of time where the source was published or last modified.
8. **Type**: the nature or genre of the source (text, image, audio…).
9. **Format**: the file format, physical medium, or dimensions of the source (DOC, GIF, MP3…).
10. **Identifier**: the URL of the evaluated subsection or feed.
11. **Source**: a resource from which the described source is obtained, e.g. the website URL.
12. **Coverage**: the spatial or temporal coverage of the source, for instance the jurisdiction under which the resource is relevant.
13. **Language**: the language of the source.
14. **Rights**:presence of a copyright statement about the source.

The metadata-based quality measure (Appendix A) checks if each of the previous properties is available for the information source. For each individual property the scale rates from 1 (that particular property is missing) to 2 (the metadata property is available for the information source). The value for the metadata quality measure is computed by summing up the individual measurements for all the metadata properties under consideration, so the maximum value for 14 properties is 28, as shown in the following formula:

$$Metadata\ Quality(source^j) = \sum_{i=1}^{14} Presence(property_i, metadata^j)$$

$$Presence(property_i, metadata^j) = \begin{cases} 1, & property_i \in metadata^j \\ 2, & property_i \notin metadata^j \end{cases}$$

The IQm also includes a content-based quality measure (Appendix B). It consists in a scale rate from 0 to 3, where 0 corresponds to the worst content quality and 3 to the best quality as detailed in Appendix B, measured for 10 parameters that focus on the following content attributes, also adopted from previous EFSA projects:

1. **Accessibility**: physical conditions in which users can retrieve the source content.
2. **Relevance**: refers to whether the source provides relevant information.

3. **Accuracy**: accurate information sources provide a reliable and valid representation of reality.
4. **Edition**: raw or processed (maximum quality).
5. **Timeliness**: the amount of time between when an event occurs and when an information source reporting about it is made available. This parameter will be measured once monitoring has started, currently it has not been measured.
6. **Clarity**: whether the information source is accompanied by appropriate metadata. Basic metadata must be provided on species affected, cause, what was observed, etc.
7. **Comparability**: the amount of available information compared to the amount that was originally expected. If data presented is sufficiently complete to be proposed as part of an umbrella for an emerging plant health threat.
8. **Coherence**: whether the information source uses recognised standards for content items.
9. **Authority**: the amount of information available about the information source.
10. **Reputation**: the impact of and information source in terms of citations in the context of a scientific community. This indicator is only applied to scientific publications published in journals with an impact factor, like Thomson Reuters Journal Citation Reports®.

The previous content attributes are measured individually and then combined using the following formula to compute content quality:

$$Content\ Quality(source^j) = \sum_{i=1}^{10} Attribute_i(source^j)$$

The measures for each of the 10 content attributes range from 0 to 3, so the maximum content quality measure for a source will be 30.

Finally, each information quality parameter is quantified using 2x2 IQ metrics (see Appendix C), to rate each source with a unique numeric value and see the differences in information quality between different types of collected sources. The results of applying this approach are reported in Section 3.1.1.

2.2.2. Indirect Method: Web Search-based Selection of Media Sources

Besides the manual selection of already known sources, the project applied an alternative approach to identify unknown sources that could be monitored with MedISys. These sources help the early detection of re-emerging and especially new emerging plant health threats because they mainly correspond to blogs and other social media sites, which have already been identified as a valuable source for early warning information (EFSA, 2012).

The approach is based on searches for different keywords in search systems available in the Web. The Web search systems used are:

- Bing: Microsoft's search engine that provides services for web search and news search.
- Feedzilla: a free RSS news service.
- Faroo News Search: includes news articles from newspapers, magazines and blogs.

All these search engines are free to use (for a limited number of requests) and provide an API, allowing automating the search process. It is important to notice that Google's Terms of Service do not allow the use of automated queries of any sort. Therefore, this search system was not considered.

Identification of appropriate keywords

In order to search for relevant sources, it is necessary to identify the appropriate keywords to use. The Plant Health Threat Ontology described in Section 2.1 relates each pest to their common,

scientific and other names, which were used as the search keywords. Therefore, for each pest defined in the ontology, a list of search keywords was obtained with the following SPARQL query:

```
PREFIX taxon: <http://purl.uniprot.org/core/>
PREFIX pests: <http://rhizomik.net/ontologies/pests#>

SELECT *
FROM <http://rhizomik.net/pests/>
WHERE {
   ?r a pests:Pest
   OPTIONAL { ?r rdfs:label ?label }
   OPTIONAL { ?r taxon:scientificName ?scientific }
   OPTIONAL { ?r taxon:otherName ?other }
   OPTIONAL { ?r taxon:commonName ?common }
   OPTIONAL { ?r taxon:host ?host }
}
```

This query returns related names for each pest, as shown in Table 5.

Table 5: Example of related names and keywords

Pest	Scientific name	Other names	Common names
Squash leaf curl virus	Squash leaf curl virus	Squash leaf curl begomovirus	SLCV, leaf curl of melon, curly mottle of watermelon,…
Bean golden mosaic virus	Bean golden mosaic virus	Bean golden mosaic geminivirus	BGMV, golden mosaic of beans, mosaïque dorée du haricot,…
Euphorbia mosaic virus	Euphorbia mosaic virus	Euphorbia mosaic geminivirus	EuMV
Phytophthora ramorum	Phytophthora ramorum	BBA 9/95 CBS 101553	Sudden oak death, Sudden oak disease, Ramorum blight, Muerte súbita del roble,…

Search process

The obtained keywords were used to perform the Web search. The search was limited to 50 results for each keyword because beyond this limit most results were usually not relevant. The previous search engines, through their APIs, returned a list of results and for each one they provided the following details:

- Webpage URL
- Title
- Description
- Position inside the list of results

Additionally, the following information was added for each result:

- **Number of matches**: the number of times that the keywords appeared in that concrete page. This information was kept and added to the results to facilitate assessing the relevance of each result because it is not the same that they mention a pest once or ten times.
- **Content language**: the language of the page content, because it was a field required to describe the information sources included in deliverable ES1.

- **Last accessed date**: the last time that page was accessed, to keep track of the last time an information source was returned by a query.

Table 6 shows an example of result obtained from Web search. These data are used to describe the information sources as well as to filter out non-relevant sources. The URL and Content Language properties were used for describing the sources as required in deliverable ES1. The Title, Description, Position and Number of Matches properties were used to review sources and prioritise them. Although the Web search was an automated process, it was necessary to manually review the resulting sources and discard those that were not relevant, mainly because the featured keywords or part of them were ambiguous terms also used in domains not related to plant health. Finally, there is the Last Access date, which describes when was the last time that the source was retrieved.

Table 6: Example of data obtained for a Web search result

Property	**Value**
URL	http://www.therepublic.com/w/IN--Ash-Borers
Title	Purdue insect expert: Emerald ash borers likely survived Indiana's frigid winter
Description	WEST LAFAYETTE, Indiana — A Purdue University entomologist says an invasive insect that's taken a big bite out of Indiana's ash tree populations likely survived the frigid winter with few losses in its numbers…
Position	25
Number of Matches	5
Content Language	en (English)
Last Access	2014-03-27

3. Results

This section details the outcomes of the project based on the previous foundations, the Plant Health Threat Ontology and the proposed methodologies to collect the news sources to be monitored by MedISys. First, there are the results about the proposed sources to be monitored for both approaches, direct and indirect. Then, the MedISys categories performing the news items selection are presented. They include categories for known threats based on the multilingual source of threat names captured by the ontology. There are also categories for unknown threats. In this case, two different approaches were tested. First of all, categories based on plant health experts that manually curated the terms to be monitored. Second, an approach based on the ontology that monitors, for some pests, terms associated to them but avoids their names. This way, terms associated to vectors, crops or symptoms are monitored. This section concludes with evaluations of the results produced by the generated MedISys categories and how they are reported through MedISys user interface.

3.1. Collection of Sources for MedISys Monitoring

The process of sources collection based on both the indirect and direct approaches continued during the project and resulted in a set of 1945 sources to be monitored, as summarised in Table 7. For the direct approach, the main improvement compared with what was available in the first interim report IR1 was the addition of some journals from the PestLens list that satisfied MedISys requirements for their monitoring. They are now available from "ES1 - MedISys Sources.xlsx", which contains 61 direct sources available from rows 2 to 62.

For the indirect approach, as it is based on searching through Bing using keywords from the ontology, and it was largely enriched with more keywords in more languages, the process generated a big amount of potential results that were evaluated and added to the updated version of ES1. As a result of this source gathering and evaluation process, ES1 was updated from the previous 618 indirect sources, those evaluated till September 2014. With this last update, the amount of indirect sources increased to reach 1884 sources, available in "ES1 - MedISys Sources.xslx" from rows 63 to 1946.

Additionally, though they have not yet been considered for monitoring by MedISys, 311 journals were analysed for their potential monitoring by MedISys in case this system is adapted to monitor academic journals. The full list of journals is available also from "ES1 - MedISys Sources.xlsx".

Table 7: Summary of information sources reviewed, evaluated and finally selected to be monitored

	Selected Sources
Direct Method	61
Indirect Method	1884
Total	1945

3.1.1. Direct Method Results

A set of 1028 known sources coming from 6 different collections was reviewed as detailed in Table 8. This included browsing known sources web pages, looking for alert systems or news feeds in these sources, identifying interesting subsections in the corresponding websites, etc. It is also important to note that the final list of sources to monitor includes only media sources as it was agreed with EFSA, so other kinds of sources are discarded.

Table 8: Known information sources reviewed using the direct method

Collections of known sources	Provided by	Sources reviewed	Notes about the review and evaluation process	Sources evaluated[5]
Sources proposed in the project call	IRTA	103	Only sources related to the plant health domain have been evaluated. Databases, grey literature, general search engines, etc. have been excluded. The Dataquest (EFSA) project sources have been included in this set.	29
Document Project nº 185e. Annex A (see page 31).	EFSA	188	Evaluation: from 188 sources, only those related to the project have been evaluated. 51 sources are related to pests.	33
Document Project nº140.	EFSA	74	Annex IV. List 2. URL: http://edepot.wur.nl/240041; EFSA call: http://edepot.wur.nl/240041	31
Document Project: PRASSIS	EFSA	533	Evaluation: only sources related to the project have been evaluated.	6
EPPO	EPPO website	63	Evaluation: EPPO members website is reviewed in order to include them as potential interesting official sources. They are National Plant Protection Organizations (NPPOs). This is a work in process because not all the necessary information is available in this website. The solution is to consult external sources such as government websites and find the plant health section or department.	21
PestLens ("Other" tab in the spread sheet)	JRC	67	32 sources were already selected from other collections. 3 repeated sources in the PestLens list. 1 source is missing the corresponding Web link and it has been impossible to identify it in the Web. 12 sources were not relevant because their main topic or because they did not provide a news feed. 19 were finally evaluated.	19
Totals		**1028**		**139**

To summarize, 1028 sources were reviewed and (after discarding repeated, non-relevant or not suitable for monitoring sources) 139 were filtered to be further evaluated using information quality measures, as detailed in the accompanying document "EvaluationSources_DirectProcess.xlsx".

Based on these quality measures, 61 were finally selected as information sources to be monitored by MedISys. All these sources, whose identifier contains letter "D", are listed in the deliverable ES1.

One of the factors causing the big reduction in the amount of sources under consideration as a result of the review and evaluation process, from 1028 to 61 sources, is due to the fact that many of them are static and thus cannot be monitored by MedISys. Although they might contain relevant information, they do not feature alert systems or news sections that can be monitored. Table 9 shows some examples of information sources that were discarded due to this fact.

Table 9: Examples of information sources discarded because they are static and thus cannot be monitored

Type of Source	Source name
Official	• Plant Pest Surveillance. Canadian Food Inspection Agency (CFIA) http://www.inspection.gc.ca/eng/1297964599443/1297965645317

[5] The file "EvaluationSources_DirectProcess.xlsx" provides details about the selection process and criteria used for the selection for each individual evaluated source

	• NAPIS Pest Tracker http://ceris.purdue.edu/ceris/ • Canadian Food Inspection Agency Plant Pest Surveillance http://www.inspection.gc.ca • The Food and Environment Research Agency (FERA) Plant Health section: UK Plant Health Risk Register http://www.fera.defra.gov.uk/ • Integrated Plant Health Information System (IPHIS) http://www.aphis.usda.gov/wps/portal/aphis/home • The Danish AgriFish Agency Quarantine Pests Online Section, Search Mechanism http://agrifish.dk/
Scientific or Research	• Greenpeace http://www.greenpeace.org/international/en/ • Utah Pests News Quarterly Newsletter http://utahpests.usu.edu • Hawaii Early Detection Network Invasive Species Lists of Hawaii http://dlnr.hawaii.gov/hisc/ • Invasive Species South Africa (ISSA) Newsletter http://www.invasives.org.za • Bugwood Blog http://www.invasive.org/ • International Phytoplasmologist Working Group http://www.ipwgnet.org/

As expected, the known information sources considered by the direct method obtained a high quality value. In fact, 90% of the sources received a score above 43,5 points, which is the threshold in IQm above which an information source is considered a relevant source. The maximum quality is 57, as shown in Table 10. Moreover, the Reputation attribute applies just to journals so for the rest of sources the maximum quality is 54.

Table 10: Source assessment parameters (2x2 IQ metrics)

Metadata Quality Properties	**Maximum Score**
Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source Coverage, Language and Rights.	30
Data Quality Attributes	
Accessibility, Relevance, Accuracy, Edition, Clarity, Comparability, Coherence, Authority and Reputation.	27
Total Score (2x2 IQ metrics)	**57**

The deliverable ES1 lists all the selected information sources using the direct method. They can be distinguished from the sources selected using the indirect method, described next, because their identifier starts with the letter "D". These are the sources whose quality measure is above the 43,5 threshold.

Most of the obtained information sources (about 67%) correspond to scientific and research data as shown in Figure 6. Official information sources represent the 21% and other media sources represent

the remaining 12%. Official information sources received the highest scores, followed by scientific and research sources.

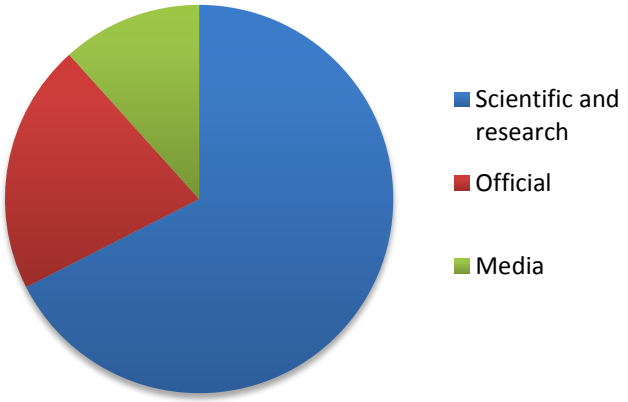


Figure 6: Distribution of selected information sources by type

3.1.2. Indirect Method Results

The Indirect Method was performed from March 2014 to July 2015 during almost 18 months. The monthly Web searches generated 2776 results. The search APIs that were employed already restrict results to newspapers, blogs and other periodic media so MedISys can monitor most of these sources. Consequently, all 2776 where evaluated. Each of them was evaluated using the Content Quality Relevance attribute, i.e. whether or not content was really related to any pest from the ontology. Following this measure, 1884 were selected as relevant to monitor as detailed in the accompanying file "EvaluationSources_IndirectProcess.xlsx".

Table 11: Summary of information sources reviewed, evaluated and finally selected to be monitored

	Reviewed Sources	**Evaluated Sources**	**Selected Sources** (to be monitored)
Indirect Method	2776	2776	1884

The complete information quality management process IQm was not applied to the information sources collected using the indirect method, mainly due to the high volume of results and because search engines already prioritise results using similar criteria. However, each retrieved source was evaluated at least using the content quality Relevance attribute. It was checked if the content of the information source was related to any pest from the ontology. Furthermore, the tool recorded the number of news items retrieved from the same information source as a new quality parameter that was used to prioritise the information sources that were frequently collected by the indirect method.

3.2. MedISys Categories

This subsection describes all the categories generated to guide MedISys monitoring. They include categories for named threats, useful when the news items explicitly mention the plant pest or disease, but also two different approaches to generate categories that do not use threat names. First, those based on a manually curated combination of terms associated to new or unknown threats defined by plant health experts. Second, an experimental approach generating categories using terms for concepts associated to a selected set of pests and diseases as modelled in the Plant Health Threat Ontology. Concretely, affected crops, vectors and symptoms.

3.2.1. Named Threats Categories

One of the objectives of the project was to monitor news for known threats, including existing threats geographical spread and re-emerging threats. In these cases, the most accurate approach is to use plant health pests or diseases names as the basis for MedISys categories. Accordingly, a list of pest and diseases to be monitored was agreed and modelled in the ontology as described in Section 2.1.2. The reported 117 threats were modelled in the ontology including different kinds of names (scientific, common, …) which were then used to generate the MedISys categories for named threats. The approach generated a category for each pest or disease, which includes a weighted word list that features all the names associated with it, as detailed in Table 12. Each one of them has a weight equal to the category threshold, consequently, just one of them is enough for the corresponding category to select a news item.

Table 12: Approach followed to generate MedISys named threats categories from the ontology

Weighted Words List	Threshold 100 - Scientific names: 100 - Common names: 100 - Other names: 100

Overall, **1609** labels were included for the 117 threats. They are listed per language in Table 13. The categories for the initial set of 47 pests and diseases are included in the accompanying file "ES2 - MedISysCategories-EFSAPlantHealthOntology.xlsx" and the additional 70 added till the end of the project are available from "ES2 - MedISysCategories-EFSAPlantHealthOntology-Ext.xlsx".

Table 13: Number of labels included in the ontology for the 117 named threats per language

Language (if available)	Count	Language (if available)	Count
Not available	617	Malayalam - ml	5
Latin - la	375	Korean – ko	5
English – en	262	Danish - da	4
French - fr	81	Catalan -ca	4
German - de	68	Polish - pl	3
Spanish - es	65	Czech – cs	3
Japanese – ja	21	Hebrew – iw	3
Dutch – nl	17	Norwegian – no	3
Italian – it	16	Thai – th	3
Portuguese – pt	15	Persian – fa	2
Swedish – sv	8	Esperanto – eo	2
Finnish - fi	8	Turkish - tr	2
Chinese - zh	7	Tamil – ta	2

Russian - ru	6	Arabic - ar	2
		Total	**1609**

Evaluation

A detailed analysis of the categories based on threat names was conducted for the 47 categories listed in Table 1 and the period from February 17th to September 11th 2015. This is the longest period for which JRC provided detailed data for more than one category. The results show that just some plant health threats become mainstream, hardly 7 threats have been able to get more than 100 hits and really just *Xylella fastidiosa*, with 2717 hits, became mainstream. From the 47 categories under consideration, just 27 got at least one hit and half of them, 14, less than 10 hits.

Table 14: Number of news items selected by the 47 named threat categories in Table 1 and one or more items selected from February 17th to September 11th 2015

Threat Category	Type	Hits
XylellaFastidiosa	Bacteria	2717
RhynchophorusFerrugineus	Insects	496
AgrilusPlanipennis	Insects	356
Pomacea	Molluscs	302
HymenoscyphusFraxineus	Fungi	160
BactroceraTryoni	Insects	154
AnoplophoraGlabripennis	Insects	150
DiabroticaVirgifera	Insects	113
BursaphelenchusXylophilus	Nematodes	69
SpodopteraFrugiperda	Insects	60
CeratocystisFagacearum	Fungi	45
PhytophthoraRamorum	Oomycetes	36
TilletiaIndica	Fungi	19
SpodopteraLitura	Insects	9
TrichilogasterAcaciaelongifoliae	Insects	8
ThripsPalmi	Insects	7
GeosmithiaMorbida	Fungi	6
MoniliniaFructicola	Fungi	4
TobaccoRingspotVirus	Virus	4
PotatoSpindleTuberViroid	Virus	3
AnastrephaLudens	Insects	2
RhagoletisMendax	Insects	2
AgrilusCoxalisAuroguttatus	Insects	1
RhagoletisSuavis	Insects	1
SpodopteraEridania	Insects	1
CowpeaMildMottleVirus	Virus	1
AndeanPotatoLatentVirus	Virus	0

AndeanPotatoMottleVirus	Virus	0
AnomalaOrientalis	Insects	0
CandidatusLiberibacter	Bacteria	0
DiplocarponMali	Fungi	0
EuphorbiaMosaicVirus	Virus	0
HeterobasidionIrregulare	Fungi	0
LettuceInfectiousYellowsVirus	Virus	0
NacobbusAberrans	Nematoda	0
PeachRosetteMosaicVirus	Virus	0
PepperMildTigreVirus	Virus	0
PotatoBlackRingspotVirus	Virus	0
PotatoVirusT	Virus	0
PunctoderaChalcoensis	Nematoda	0
RhagoletisCingulata	Insects	0
RhagoletisFausta	Insects	0
RhagoletisIndifferens	Insects	0
RhagoletisRibicola	Insects	0
StrawberryVeinBandingVirus	Virus	0
TeciaSolanivora	Insects	0
ThecaphoraSolani	Fungi	0

This was even more evident at the end of the project, when JRC provided detailed data for just one category, the most active one, *Xylella fastidiosa*. The data for the period February 17th 2015 to June 29th 2016 shows the high activity around this threat, concretely 5082 hits just by this category. The news items selected by this category come for a very heterogeneous set of sources, concretely 712 different news sources from 77 different countries. This illustrates the multilingual capabilities of the ontology and the categories generated from it. Table 15 details the number of news items selected for news sources from the 10 most active countries. As it can be observed, more than half of the hits, 3198, come from Italian news sources.

Table 15: Number of of news items selected by the *Xylella fastidiosa* category from February 17th to June 29th 2016 for the 10 most active countries the corresponding news source corresponds to

Country	Hits
Italy	3198
France	608
Spain	321
Greece	128
Belgium	121
USA	118
Germany	71
Great Britain	68
Luxembourg	43
Netherlands	28

A detailed review of the news items selected by the named threats categories was carried out by following an iterative approach. First of all, a selection of 5 categories was created in order to test them. The selected pests were those more present in results obtained from the web search-based selection of media sources. They were: *Agrilus planipennis*, *Anoplophora glabripennis*, *Bactrocera tryoni*, *Phytophthora ramorum*, and *Xylella fastidiosa*.

The first version of these categories included the word weight list with names of the pest and a basic set of combinations in English. After two weeks, selected articles for each category were reviewed. On the one hand, many articles were correctly selected because they mentioned one of the pest names thanks to the words weighted list. On the other hand, some non-related articles were selected because they matched one of the combinations.

In order to reduce the noise three main improvements were proposed and implemented for the rest of the project:

- **Introduce proximity in combinations**: a combination without proximity is not effective and it selects unrelated articles which contain the terms but without any relationship, in different paragraphs, etc. Regarding this matter, JRC suggested to use a low proximity value when the combination must contain two terms and a bigger value when the combination contains more terms. The current proposal is to use proximity 15 in combinations of two terms and 100 in combinations of three or more terms.
- Avoid the usage of '%' operator, which produces a lot of noise. Instead, it seems more appropriate to add related terms such as singular and plural forms, verb conjugations, etc.

Additionally, a larger scale test was later conducted for the 47 categories listed in Table 1 and using just words weighted lists containing scientific, other scientific and common names, though in all available languages. The results were also very satisfactory. 100 news items were checked manually and just for one threat non-relevant results were identified. This was "Pomacea" and the problem was with one of its common names in Portuguese: "Aruá". MedISys required that special characters like accents are replaced with the wildcard "_" so even if the word is not properly written or there is an encoding error, the corresponding information source can be selected by the category.

However, when words are too short, like in this case, this might make the category too generic and then select documents containing unrelated words like "Aruk" or "Arus". To solve this issue, the final versions of the automatic category generation tool detects these cases. When the term is too short it avoids generating the version with the wildcard. However, for words with accents, the approach is to generate two versions, one with accent and the other without, i.e. to generate "Arua" and "Aruá". This way we minimise the news items missed due to this simplification because usually the approach is simply to omit the accent.

The procedure for the analysis of the news articles is detailed next to facilitate its reproducibility:

1) Analysis of the news articles catched by Medisys was done selecting the news, placing them into the 'Newsletter', and then exporting it to XML. An XML editor (XML Notepad 2007) allows manually selecting the relevant information for each news item (title, URL, trigger words, etc.). This allows keeping record of the obtained links, trigger words, as well as the other categories that trigger each alert, as shown in Table 16.

Table 16: XML metadata sample obtained after exporting from MedISys NewsDesk a particular news item selected by a category

```
<item>
        <title>Natural Resources Board expands list of invasive species</title>
        <link>http://www.jsonline.com/news/wisconsin/natural-resources-board-expands-list-of-
invasive-        species-b99406748z1-285388931.html</link>
        <description>. Journal Sentinel files The state Natural Resources Board on Wednesday
downgraded       the invasive emerald ash borer from prohibited to restricted. The step acknowledges
that the         insect is here, it's spreading and       it's not likely that the ash-killing insect
will ever be     eradicated.</description>
        <pubDate>2014-12-10T20:19+0100</pubDate>
        <source url="http://www.jsonline.com/rss/?c=y&c=y&path=%2F&path=%252F"
        country="US">jsonline</source>
        <category emm:trigger="emerald ash borer[5]; ">oak_EPPO_1_A_1</category>
        ...
</item>
```

2) In order to analyse the contents of each alert (link) and improve the query (i.e. understand why it was selected when clearly not related to any plant health threat), each URL has to be opened, and the trigger word were searched in the text. Initially, this was a slow process because each word was looked separately. The recommendation is to use of a browser extension to search multiple words (e.g. the SearchWP https://code.google.com/p/searchwp/ plugin for Firefox) facilitates this work as it converts the search box into a dynamic tool that reveals the text we are looking for, i.e. the trigger words, with colour-coded highlighting, as shown in Figure 7. This way it was possible to test for each item if it was properly selected and actually related to plant health using the original content of the news items that made clear the context of each trigger word that made the corresponding category select the news item.

Figure 7: Using SearchWP Firefox plugin to highlights category trigger words in the news item

3.2.2. Categories based on Manually Curated Terms for New Threats

The most direct approach to generate a MedISys category to monitor unknown threats is to generate a category based on a list of manually curated terms that usually are present in news items about plant health threats. Used keywords do not include any names of pest or disease (nor scientific, nor common like aphid, caterpillar, etc.), which are already covered by the categories generated for Objective 2. This category complements the previous ones because they are capable of selecting news items that do not explicitly mention a plant health threat using one of its names.

As detailed in

Table 17, this category is based on word combinations, with three sets of alternative words and one of negative words. In order to select a news item, it should contain at least one word for each of the 3 alternative sets and none of the words in the negative set.

Table 17: Word combinations for the manually curated category for unknown threats

combination		
	proximity	15
	or	alien, danger, dangerous, deadly, emerge, emerged, emerging, infest, infestation, infestations, infested, infests, invade%, invasion, invasive%, mysterious, new+species, outbreak, outbreaks, recent, spread, spreading, spreads, strange, unexplained, unidentified, unknown
	or	agricultural, agriculture, almond, almonds, apple, apples, apricot, apricots, arable+crop%, ash, aubergine, aubergines, barley, bean, beans, beet, beets, berries, berry, blueberries, blueberry, broccoli, cabbage, cabbages, carrot, carrots, cauliflower, cauliflowers, cereal, cereals, cherries, cherry, chesnuts, chestnut, citrus, corn, cotton, crop, crops, cucumber, cucumbers, cucurbit, cucurbits, elm, flower, flowers, forage, forest, forestry, forests, fruit, fruits, garlic, grape, grapes, hay, hazelnut, hazelnuts, horticultural, horticulture, legume, legumes, lettuce, lettuces, maize, nectarine, nectarines, oak, oaks, olive, olives, onion, onions, orchard, orchards, ornamental, ornamentals, palm, palms, pasture, pastures, pea, peach, peaches, pear, pears, peas, pepper, peppers, pine, pines, pistachio, pistachios, plant, plants, plum, plums, pome+fruit, pome+fruits, potato, potatoes, pumpkin, pumpkins, rice, root+beet, rye, soya, soybean, soybeans, spinach, stone+fruit, stone+fruits, strawberries, strawberry, sugarbeet, sugarbeets, sugarcane, sugarcanes, tomato, tomatoes, tree, trees, vegetable, vegetables, vineyard, vineyards, walnut, walnuts, wheat, wine, wines
	or	bacteria, bacterial, crop+failure, crop+loss, damage, damaged, damages, damaging, death, decline, dieback, disease, diseases, epidemic, epidemy, fungal, funghi, fungus, illness, infection, infections, injuried, injury, insect, insects, loss, mite, mites, mortalities, mortality, nematode, nematodes, pest, pests, phytoplasma, phytoplasmas, plant+health, risk, risks, sickness, threat, threatens, threats, viral, virioid, virus, yield+loss
	not	allergies, allergy, animal+abuse, beatles, beef, berlusconi, caffeine, canine, central+park, chickenpox, civil+unrest%, dementia, dog, earthquake%, ebola, facebook, factory+farm, factory+farms, fever, flames, flood, flooding, floods, food+basket, food+baskets, fukushima, gay, google, hailstorm, haistorms, hay+fever, healthy+eat%, healthy+food%, haemorrhag%, hemorrhag%, herders, humanitarian, iMac, incend%, industrial+production, iPad%, iPhone%, Lennon, lesbian, measles, mental+health, microsoft, Mr+Bean, mumps, narcotic%, nokia, nozzel, nuclear+industr%, nuclear+reactor%, pork, poultry, power+line%, sheep, sheeps, smartphone%, spreads+market, staphilococcus+aureus, storm%, suicid%, swine, train+service%, tsunami, unseasonal+rain, violence, volcano%, wall+street, wars

This category is available from the accompanying file "ES3 - MedISysCategories-NewPlantPests.xlsx" and its results from:

http://medisys.newsbrief.eu/medisys/alertedition/en/new_plant_health_threats.html

During 2014 and 2015 this category was improved by analysing results for some periods (e.g. by adding more NOT terms, and removing wildcards).

As this category triggered many not useful news, and as proposed by EFSA during project meetings, three other categories were also generated in order to see if it was possible to reduce the level of 'noise' without affecting the number of triggered news.

a) ... new_pl_pests_9: reducing the proximity to 9 (instead of 15);

b) ... new_pl_pests_9_alerts: maintains the 9 proximity, but adds another group to cross with key-words associated with emergency-related events;

c) ... new_pl_pests_A: maintains the original 15 proximity, but redistributes some of the key-words between the groups in order to obtain greater similarity ('new', 'pests', 'loss`, 'crops').

The category combinations of those alternatives are listed in Table 18. They are available from

http://medisys.newsbrief.eu/medisys/alertedition/en/oak_new_plant_pests_9.html

http://medisys.newsbrief.eu/medisys/alertedition/en/oak_new_plant_pests_9_alerts.html

Table 18: Additional alternative word combinations to identify unknown PHT.

Combination: oak_new_plant_pests_9_alerts

combination		
	proximity	9
	or	alien danger dangerous deadly emerge emerged emerging infest infestation infestations infested infests invade% invasion invasive% mysterious new+species outbreak outbreaks recent spread spreading spreads strange unexplained unidentified unknown
	or	bacteria bacterial crop+failure crop+loss damage damaged damages damaging death decline dieback disease diseases epidemic epidemy fungal funghi fungus illness infection infections injuried injury insect insects loss mite mites mortalities mortality nematode nematodes pest pests phytoplasma phytoplasmas plant+health risk risks sickness threat threatens threats viral virioid virus yield+loss
	or	alarm alert emergency crunch crisis critical severe disaster cataclysm calamity break collapse debacle deluge rush urgency
	or	agricultural agriculture almond almonds apple apples apricot apricots arable+crop% ash aubergine aubergines barley bean beans beet beets berries berry blueberries blueberry broccoli cabbage cabbages carrot carrots cauliflower cauliflowers cereal cereals cherries cherry chesnuts chestnut citrus corn cotton crop crops cucumber cucumbers cucurbit cucurbits elm flower flowers forage forest forestry forests fruit fruits garlic grape grapes hay hazelnut hazelnuts horticultural horticulture legume legumes lettuce lettuces maize nectarine nectarines oak oaks olive olives onion onions orchard orchards ornamental ornamentals palm palms pasture pastures pea peach peaches pear pears peas pepper peppers pine pines pistachio pistachios plant plants plum plums pome+fruit pome+fruits potato potatoes pumpkin pumpkins rice root+beet rye soya soybean soybeans spinach stone+fruit stone+fruits strawberries strawberry sugarbeet sugarbeets sugarcane sugarcanes tomato tomatoes tree trees vegetable vegetables vineyard vineyards walnut walnuts wheat wine wines
	not	allergies allergy animal+abuse beatles beef berlusconi caffeine canine central+park chickenpox civil+unrest% dementia dog earthquake% ebola facebook factory+farm factory+farms fever flames flood flooding floods food+basket food+baskets fukushima gay google hailstorm haistorms hay+fever healthy+eat% healthy+food% haemorrhag% hemorrhag% herders humanitarian iMac incend% industrial+production iPad% iPhone% Lennon lesbian measles mental+health microsoft Mr+Bean mumps narcotic% nokia nozzel nuclear+industr% nuclear+reactor% pork poultry power+line% sheep sheeps smartphone% spreads+market staphilococcus+aureus storm% suicid% swine train+service% tsunami unseasonal+rain violence volcano% wall+street wars

Combination: oak_new_pl_pests_A

combination		
	proximity	15
	or	alien emerge emerged emerging invade% invasion invasive% mysterious new+species outbreak outbreaks recent spread spreading spreads strange unexplained unidentified unknown
	or	crop+failure crop+loss damage damaged damages damaging danger dangerous deadly death decline dieback epidemic epidemy infest infestation infestations infested infests injuried injury loss mortalities mortality risk risks sickness threat threatens threats yield+loss
	or	bacteria bacterial disease diseases fungal funghi fungus illness infection infections insect insects mite mites nematode nematodes pest pests phytoplasma phytoplasmas plant+health viral virioid virus
		agricultural agriculture almond almonds apple apples apricot apricots arable+crop% ash aubergine aubergines barley bean beans beet beets berries berry blueberries blueberry broccoli cabbage cabbages carrot carrots cauliflower cauliflowers cereal cereals cherries cherry chesnuts chestnut citrus corn cotton crop crops cucumber cucumbers cucurbit cucurbits elm flower flowers forage forest forestry forests fruit fruits garlic grape grapes hay hazelnut hazelnuts horticultural horticulture legume legumes lettuce lettuces maize nectarine nectarines oak oaks olive olives onion onions orchard orchards ornamental ornamentals palm palms pasture pastures pea peach peaches pear pears peas pepper peppers pine pines pistachio pistachios plant plants plum plums pome+fruit pome+fruits potato potatoes pumpkin pumpkins rice root+beet rye soya soybean soybeans spinach stone+fruit stone+fruits strawberries strawberry sugarbeet sugarbeets sugarcane sugarcanes tomato

		tomatoes tree trees vegetable vegetables vineyard vineyards walnut walnuts wheat wine wines
	not	allergies allergy animal+abuse beatles beef berlusconi caffeine canine central+park chickenpox civil+unrest% dementia dog earthquake% ebola facebook factory+farm factory+farms fever flames flood flooding floods food+basket food+baskets fukushima gay google haemorrhag% hailstorm haistorms hay+fever healthy+eat% healthy+food% hemorrhag% herders humanitarian iMac incend% industrial+production iPad% iPhone% Lennon lesbian measles mental+health microsoft Mr+Bean mumps narcotic% nokia nozzel nuclear+industr% nuclear+reactor% pork poultry power+line% sheep sheeps smartphone% spreads+market staphilococcus+aureus storm% suicid% swine train+service% tsunami unseasonal+rain violence volcano% wall+street wars

Evaluation

A sample of 100 results for the categories based on manually curated terms and targeting unknown threats was analysed during the project. The study followed the same approach detailed in Section 3.2.1 for named threats categories. The selected news items for the categories were selected in MedISys NewsDesk and then exported to XML to be able to analyse their associated metadata. Then, they were manually inspected in the context of the news item context using a plugin that highlighted all trigger words. The review showed that 78 out of 100 items were actually related with plant health and thus relevant based on the knowledge of the plant health expert conducting the review. The sample below corresponds to what might be expected on a daily basis. From these four items, the first three are related to plant health while the last one is unrelated but uses terms from the category. For each one, all the available metadata are provided, including trigger words and the categories that selected the news item.

Kenya: El Niño Won't Hit Food Costs, Says Ministry

allafrica Monday, September 7, 2015 10:48:00 AM CEST | info [other]

Trigger words: [oak_new_plant_pests_15] Agriculture[3]; plant[1]; wheat[1]; outbreak[1]; rice[1]; Cereals[1]; maize[6]; beans[1]; potatoes[1]; Bean[1]; diseases[2]; agricultural[1]; crop[1];

The government has moved to allay fears of post-harvest losses as the country braces for El Niño rains. Last week, acting Agriculture Cabinet Secretary Adan Mohamed said all concerned State agencies would be meeting to finalise plans to ensure that food prices don't get out of reach for the majority of Kenyans....

Alien plants strangle local ones

thehindu Monday, September 7, 2015 9:46:00 AM CEST | info [other]

Trigger words: [oak_new_plant_pests_15] Forest[3]; plant[3]; forests[1]; tree[2]; spreading[1]; threat[1]; plants[4]; Plants[1]; trees[1]; alien[1]; forest[4]; invasive[3]; Alien[1]; spread[1];

The rampant growth of invasive alien plants is a concern for the wildlife managers in the district. “The spread of invasive plants, especially Senna spectabilis , is posing a major threat to the forest areas of the district, due to its quick growth and coppicing character,” says S. Mohanan Pillai, wildlife warden, WSS....

Biologists Climb Massive Sequoias to Gauge Health Amid Drought

nbcnews Sunday, September 6, 2015 3:48:00 AM CEST | info [other]

Trigger words: [oak_new_plant_pests_15] pines[2]; tree[1]; threat[2]; infestations[1]; trees[8]; forest[5]; insect[1]; infestation[1];

Thousands flock to the Sequoia National Park each year to view the majestic trees that tower high above the forest. But below, on the forest floor, there are signs of struggle as many Sequoia trees are shedding more leaves and foliage during the fourth year California's devastating drought....

In Alaska, you can bike and hike Kennecott's abandoned mines

sltrib Sunday, September 6, 2015 12:18:00 AM CEST | info [other]

Trigger words: [oak_new_plant_pests_15] plant[1]; infections[1]; dangerous[1];

Fairbanks, Alaska • Tires flung mud in our eyes and rain soaked every layer of clothing. The descent made our brake rotors too hot to touch, the metal sizzling in the wet conditions. Rapidly descending 4,000 feet through a rainstorm capped off the weekend of mountain biking inside Wrangell-St. Elias National Park....

This kind of analysis was carried along the whole project and allowed us to improve the list of negative words that make it possible to reduce the noise by not selecting news associated to topics unrelated to plant health. This way, terms associated to human health, cooking or sports were included in the list of negative words, as detailed in Table 17. This allowed us to reduce the amount of irrelevant news items.

The previous results are for the generic category based on manually curated terms. The additional category combining emergency-related terms and described in Section 3.2.2 was also evaluated. As requested by EFSA, this approach made it possible to reduce the volume of items selected by the generic category, about 10 per day, to about 1 per week. The 100 items analysed at the end of the project also included items selected by the emergency-related category, as is constitutes a subset of the generic one.

Just 3 items out of the 100 analysed matched this more restrictive category. This amount might be expected on a monthly basis. In the case of the evaluated set, all 3 items were relevant because they were about plant health. Even more interesting and valuable, they were not captured by any of the know threat categories so they would have been missed if the manually curated categories had not been set.

Adama's Wide Product Range Protects Potato Crops Now and in the Future

agropages Monday, August 31, 2015 6:59:00 PM CEST | info [other]

Trigger words: [oak_new_plant_pests_9_alerts] plant[2]; Agriculture[1]; fungal[1]; rice[1]; potatoes[12]; plants[1]; virus[1]; damage[2]; Potato[1]; insects[2]; disease[2]; insect[1]; diseases[2]; Crops[1]; crop[11]; break[1]; infested[1]; nematodes[2]; spreading[1]; pest[2]; crops[7]; potato[25]; severe[2]; Nematodes[1];

Entities: Van Kampen[1];

Other categories: New Plant Pests; FAO; Fungicides; Herbicides; Insecticides; Multiple Species; Pathogens;Production; Technology;

Farmers face challenges at every stage of the potato growth and storage cycle. The most well-known disease, Late Blight, estimated to cause $5.6 billion of losses per year globally,was responsible for the widespread famines in northern Europe in the 1840s that resulted in the deaths of over a million people....

Green light of hope to overcome Striga -triggered food insecurity in Africa - Fluorescent turn-on probe identifies the 'wake-up protein' in witchweed

seedquest Friday, August 21, 2015 11:12:00 PM CEST | info [other]

Trigger words: [oak_new_plant_pests_9_alerts] plant[20]; emerged[1]; infest[1]; rice[1]; threat[2]; plants[5]; Plants[1]; crops[4]; crop failure[1]; infests[1]; severe[1]; flowers[1]; corn[1]; crop[8]; infestation[1];

Other categories: New Plant Pests;

August 21, 2015. Source: Institute of Transformative Bio-Molecules (ITbM), Nagoya University A molecular approach has been used to identify the protein responsible for germination of Striga seeds through visualization by green fluorescence. Striga , a parasitic plant known as witchweed has.....

Fluorescent turn-on probe identifies the 'wake-up protein' in witchweed seeds

phys Friday, August 21, 2015 10:21:00 PM CEST | info [other]

Trigger words: [oak_new_plant_pests_9_alerts] plant[19]; emerged[1]; infest[1]; rice[1]; threat[2]; plants[5]; Plants[1]; crops[5]; crop failure[1]; infests[2]; severe[1]; flowers[1]; corn[1]; crop[8]; infestation[1];

Other categories: New Plant Pests;

Striga infests crops by absorbing nutrients and water from their roots. Credit: ITbM, Nagoya University. A molecular approach has been used to identify the protein responsible for germination of Striga seeds through visualization by green fluorescence....

The analysis of the efficacy of the *new_plant_health_threats* category to identify unknown PHT was also done with the 2015 and 2016 data files provided by JRC: 6030 records for 2015 and 3464 records for 2016 (until the 10th of July).

In order to compare the usefulness of this category (previously oak_new_plant_pests in http://medisys.newsbrief.eu/medisys/alertedition/en/oak_new_plant_pests.html) and the 3 modifications (Tables 17 and 18), we selected news that mentioned 5 European countries in the categories field of the submitted Excel files: 4 continental ones (IT FR ES DE), assuming that there would be many news concerned with new emerging PHT (e.g. *Xylella*) or because of worries about new PHT coming from eastern-EU. The UK (including Ireland) was also included because the key words are in English and the results should be more accurate than using translations from other languages.

News were considered as useful to alert for new unknown PHT or dispersal of known PHT if the text extracted in the Excel files either (a) named the PHT, or (b) used terms clearly related to diseases or pests in crops, or (c) did not directly mention a PHT, but did use very similar wording that could be related to crop losses and therefore to the identification of potential unknown PHT (e.g. reduced production or cultivation or yields or prices; crop shortages; falling or declining food prices; climate change and heatwaves; climate disrupts harvest; crop resilience; pollination problems; deforestations or loss of forest areas; emergency responses to outbreaks; mosquito plague; food borne diseases, etc.). It also included general news on PHT from EFSA, EPPO, FAO, UN, IPPC, etc. Many news were clearly duplicated one or more times in different sources and those were manually recorded separately.

Table 19 shows that the unmodified category is better, with a much higher number of triggered news. Even if there is a 40 % of non-useful news, the three modifications to this category produced a much lower number of news. Reducing proximity to 9 increased the usefulness but reduced the number of news by 45 % (2015) and 43 % (2016).

Table 19: Number of news obtained with the four categories aiming to obtain news on unspecified Plant Health Threats.

Category	number of news	usefulness (1)	number of news
	2015		2016
new_pl_pests (proximity =15)	6030	59 %	3464
new_pl_pests_A (proximity =15)	313		312

new_pl_pests_9	3286	64 %	1961
new_pl_pests_9_alerts	7		30
(1) usefulness of non-duplicated news was scored for the 5 selected countries (IT ES FR DE UK)			

Table 20 shows the detailed analysis of the news obtained with the new_pl_pests category for the 5 selected EU countries during 2015.

Table 20: Analysis of the usefulness of news obtained during 2015 with the new_pl_pests category for five selected countries mentioned in the category column.

Grouping	**IT**	**ES**	**FR**	**DE**	**UK**	**Number of news with any of those countries** (1)
useful news (original)	155	80	123	72	283	539
useful news (duplicated)	113	36	97	50	100	262
not useful news (original)	46	26	83	49	249	367
not useful news (duplicated)	21	13	18	24	82	122
TOTAL	335	155	321	195	714	1290
usefulness (only of not duplicated)	77 %	76 %	60 %	60 %	53 %	59 %
(1) Numbers do not add-up because the same new may mention more than one of those countries.						

In total, 1290 news were scored for 2015. Overall, the category produced ~60% of useful news (range between ~50% and ~75%). Therefore, for such an unspecific search for unknown or non-monitored PHT, the user should be aware of the amount of noise obtained. The use of such broad category should perhaps mainly be used by people mostly interested in identifying trends for unknown or non-monitored PHT (e.g. by EFSA, EPPO, NPPOs, and Research Institutes). Narrowing the category may lose important information, although there is still room for further improvement of the combinations of key-words to be used (e.g. Arsevska et al., 2016).

As a further indication of the usefulness of this general category (new_pl_pests) we also determined its additionality, that is to what extent the obtained news had not been obtained by any of the existing -PHT categories in MedISys (i.e. those with names for a specific PHT). Figure 8 shows the temporal trend for all news obtained during 2015 and 2016 with the new_pl_pests category, and compares them to the number of these news that were also obtained by any of the -PHT categories running at the same time in EFSA's PlantHealth section of MedISys.

All running -PHT categories also produced many other news, but they were not used here as our objective was to determine the benefits of the new_pl_pests category in identifying other news not specifically being monitored by EFSA's PlantHealth sections of MedISys.

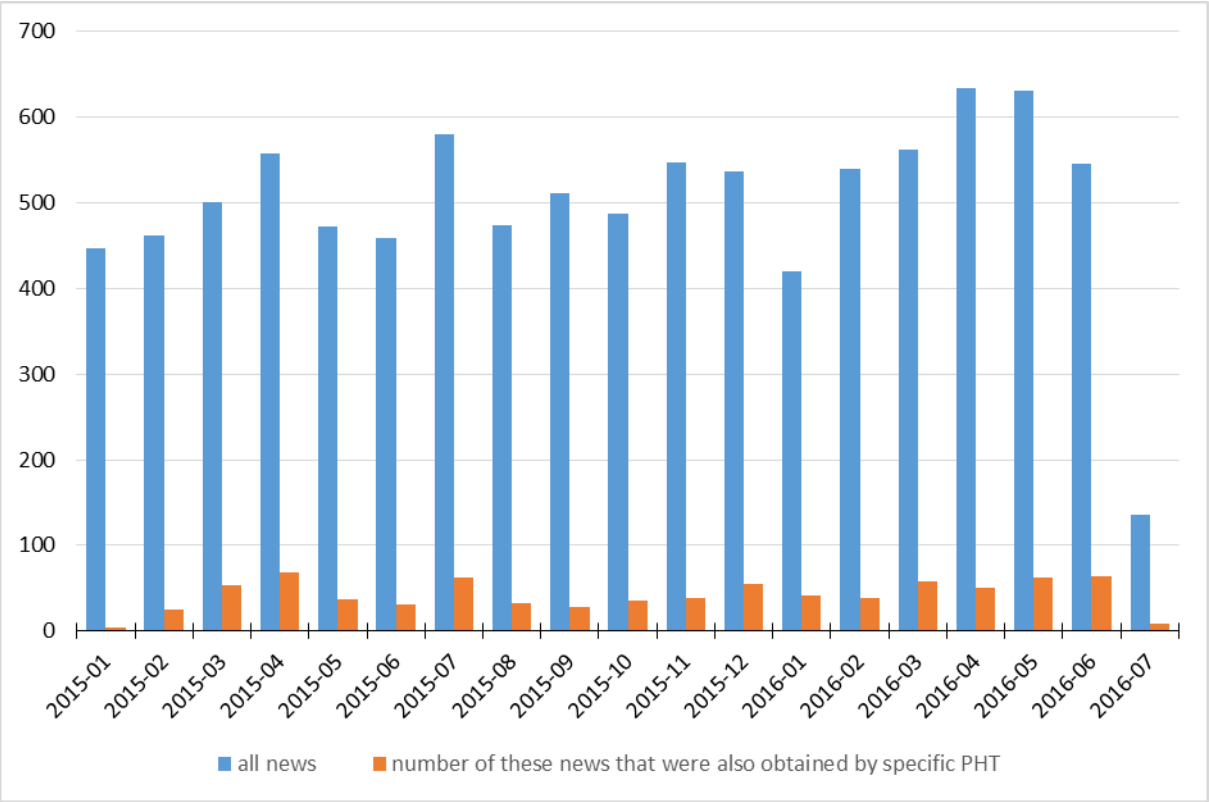


Figure 8: Number of news obtained during 2015 and 2016 with the new_pl_pests category: the blue bars show the number of all news obtained; the red bars show the number of these news that were also obtained by any of the running -PHT categories in EFSA's PlantHealth section within MedISys.

Clearly, many news were only obtained with the category for unspecified PHT (new_pl_pests) and not by any of the running specific -PHT categories in MedISys at that time. This clearly suggests its potential to identify news related to PHT that are not yet monitored by MedISys., although it requires more time to manually curate it.

Figure 9 shows a better estimate of this additionality of the category for unknown or non-monitored PHT (new_pl_pests). It shows how many of the news previously scored as 'useful' in the 5 EU countries for 2015 (Table 20) were also obtained by any of the running -PHT categories.

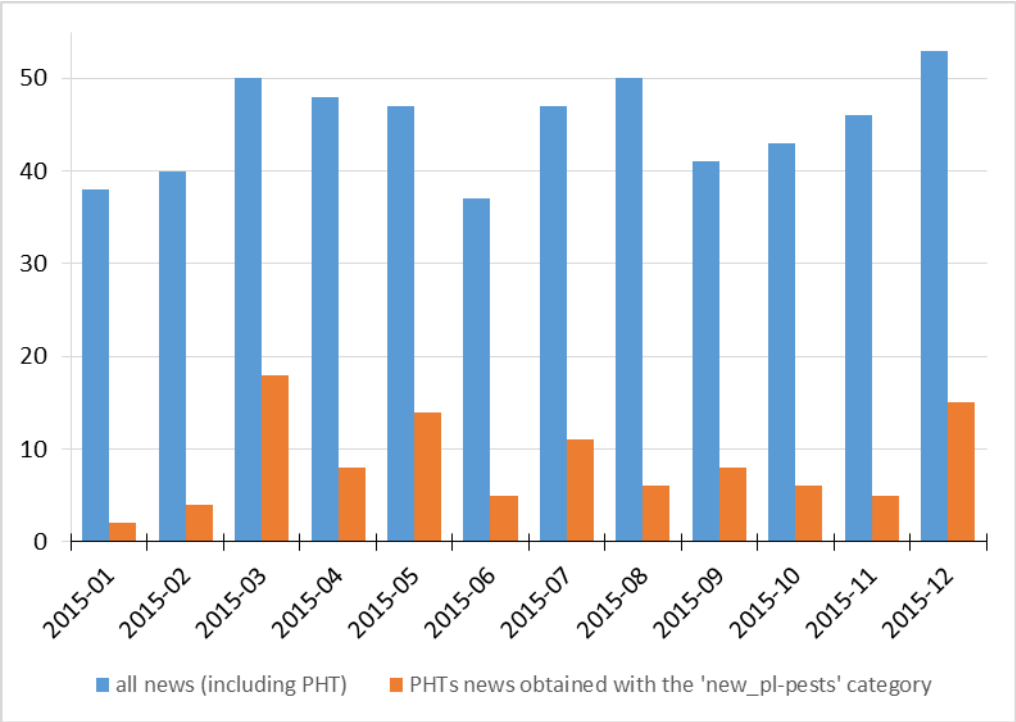
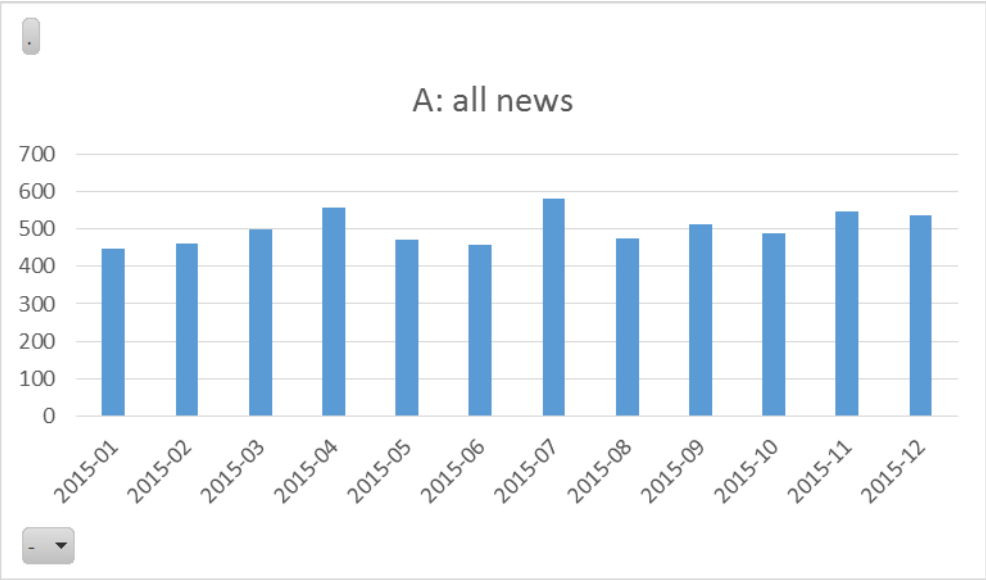


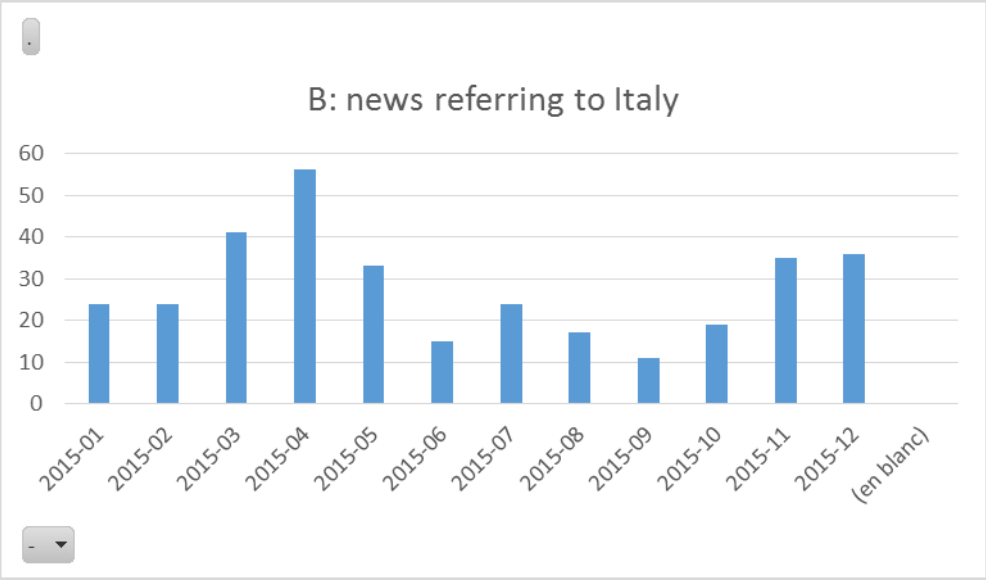
Figure 9: Number of news obtained during 2015 with the new_pl_pests category, that referred to 5 EU countries and were considered as useful after manually curating them: the blue bars show the number of all useful news obtained; the red bars show the number of these news that were also obtained by any of the running -PHT categories in EFSA's PlantHealth section within MedISys.

As before, the new_pl_pests category triggered many useful news that had not been obtained by any of the running -PHT categories. Overall, all specific -PHT categories will provide more tailored news and better information on the development of known PHT, but the category for unknown PHT (new_pl_pests) shows great potential to identify new, emerging or non-monitored PHT.

Identification of potential invasions or outbreaks: temporal trends

Figure 10 shows the monthly records obtained during 2015 for: (A) all news, (B) news referring to Italy, or (C) to Spain in the category column of the results file. Overall, when considering all the news obtained, (Figure 10A) we could not identify an increase in the number of news that would suggest more news on some PHT.





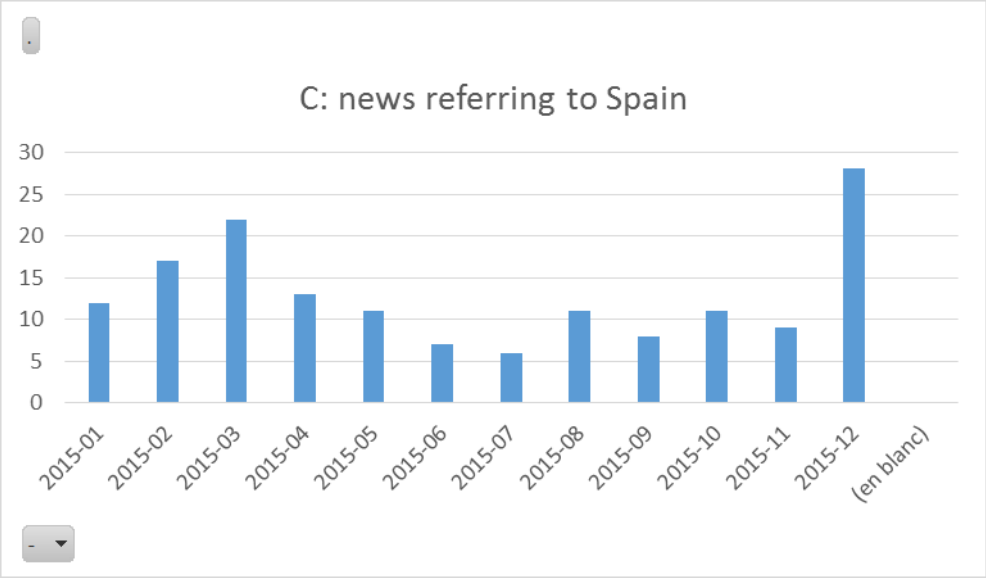


Figure 10: Temporal trend of news triggered by the new_pl_pests category during 2015. (A) 'all news', without filtering for any country; (B) news that mention 'Italy', and (C) news that mention 'Spain' in the category column of the file.

However, in the cases of 'Italy' and 'Spain' there were two increases in the number of news: in March-April for 'Italy', and in December for 'Spain'. For news mentioning 'Italy', Table 21 compares January and February vs. March and April.

Table 21: Analysis of the usefulness of news obtained with the new_pl_pests category mentioning Italy in two consecutive periods.

	2015, January and February	**2015, March and April**
useful news (original)	26	41
useful news (duplicated)	9	46
not useful news (original)	6	8
not useful news (duplicated)	7	2
TOTAL	48	97

Table 21 shows that the increase in the number of news was mostly due to useful news (from 35 to 87) and especially to the duplicated ones, thus showing the interest of media in those PHT. The use of the new_pl_pests category can therefore detect increases of news that are of interest to identify potential PHT.
To identify whether there was some specific item triggering this increase we did a deeper analysis of the 'useful' group splitting the 3 concepts mentioned previously into 3 separate sub-groups:
(a) *name* = provided the name of the PHT (scientific, popular or terms like aphid, caterpillar, fruit fly, etc.)
(b) *disease* = used terms related to diseases or pests
(c) *general* = the news used broad expressions that could also refer to crop losses

Table 22: Analysis of the contents of useful news obtained with the new_pl_pests category mentioning Italy in two consecutive periods.

useful news with 'Italy' in the categories field	***general***	***disease***	***name***	**TOTAL useful news**
2015, January and February	5	6	15 (9 on *Xylella*)	26
2015, March and April	3	8	32 (26 on *Xylella*)	41
TOTAL	8	14	47 (35 on *Xylella*)	67

As can be seen (Table 22), the increase in the number of useful news in March and April (from 26 to 41) was associated to news somehow naming the PHT (from 15 to 32), specially *Xylella*, although *Xylella* was not included in the query.

Therefore the use of a general search category for unspecified PHT seems useful to identify news mentioning new PHT. If the XylellaFastiosa-PHT category had not yet been developed, at least the use of the general new_pl_pests category would have detected an increase in the number of news referring to Italy.

The potential of the new_pl_pests category can also be seen comparing all the 2015 useful news referring to 'Italy' in the categories column with the number of these that had also been obtained by any of the -PHT categories running in MedISys at the same times (Figure 11): in all months, many additional useful news were obtained that had not been obtained by any of the running -PHT categories.

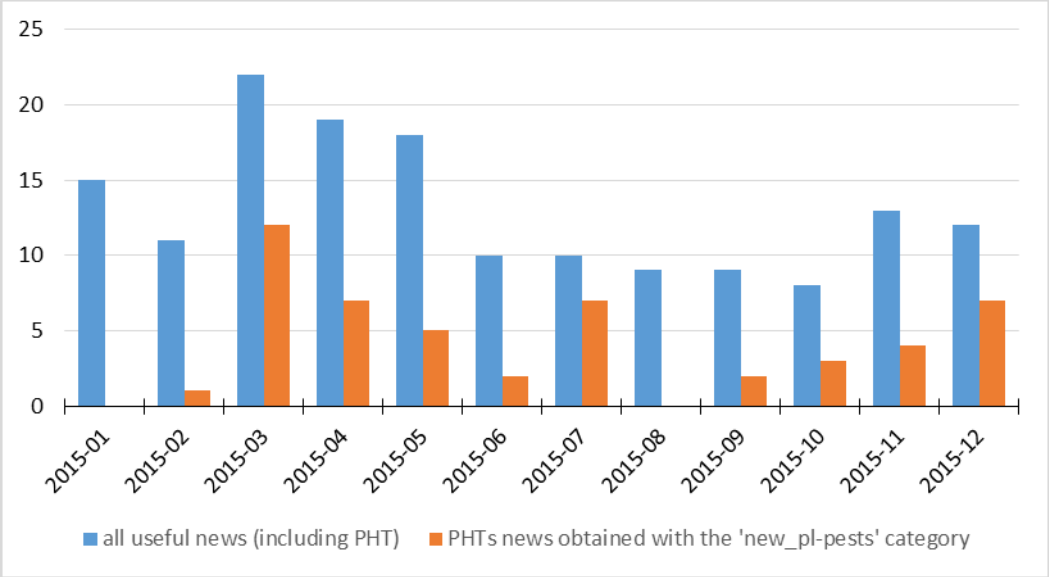


Figure 11: Number of news obtained during 2015 with the new_pl_pests category, that referred to Italy and were considered as useful after manually curating them: blue bars show the number of all news obtained; the red bars show the number of these news that were also obtained by any of the running -PHT categories in EFSA's PlantHealth section within MedISys.

For news mentioning 'Spain' in the file extracted from MedISys, there was an increase in news by December 2015 suggesting a new PHT (Figure 10C). However, from those 28 recorded news, only 11 were useful and not duplicated by other media sources, and all referred to other countries (Colombia, Australia, Ecuador, Italy, Guyana) and none specifically mentioned a worry for any PHT in Spain. Therefore, the name of countries in the categories column does not seem to aid in identifying news related to a given country. In fact, the news with 'Italy' in the categories list also included many news that were not specifically related to Italy (see a more detailed analysis in the case of UK, below).

The number of monthly news for 2015 with 'UnitedKingdom' in the categories column of the provided Excel file did not show an increase that should suggest worries about new PHT (Figure 12).

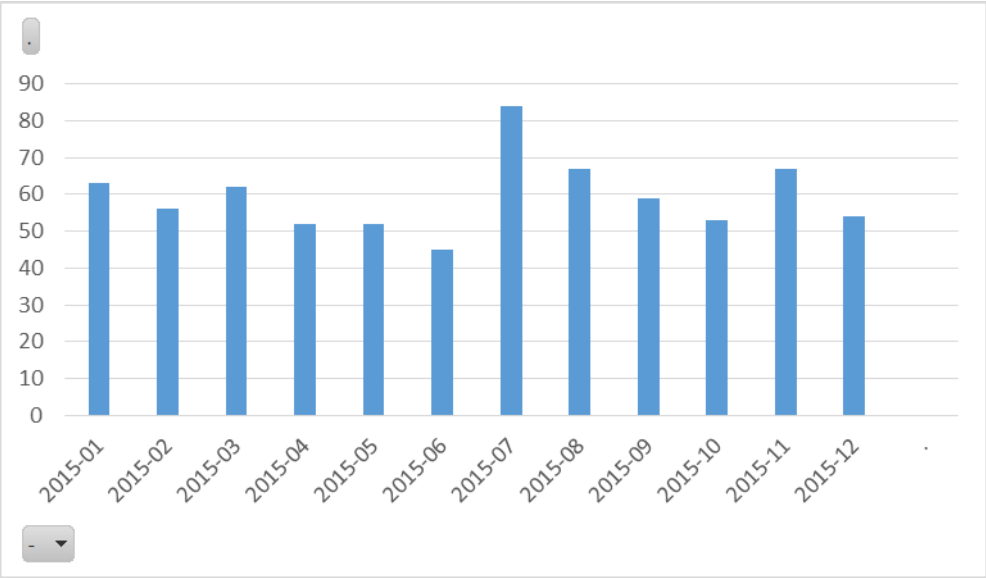


Figure 12: Temporal trend of news triggered by the new_pl_pests category during 2015 that mention 'UnitedKingdom' in the category column of the file.

We did read the text provided by MedISys in the excel file, and classified the 283 useful news (Table 20) according to the sub-groups described above. Additionally we included an 'alert ' group if the text highlighted a strong worry about a new outbreak or risk of invasion, and also classified them into yes / no according to whether they were referring to a PHT occurring or affecting the UK or not.

Table 23: Analysis of the contents of useful news obtained in 2015 with the new_pl_pests category referring to the United Kingdom.

text really refers to news about UK	***general***	***disease***	***name***	***alert***	**TOTAL**
yes	53	22	57	16	148
not	56	27	52	0	135
TOTAL	109	49	109	16	283

As Table 23 shows, almost half of the 'UnitedKingdom' news were not related to pests occurring or affecting the UK, as was also previously mentioned for Spain and Italy. Therefore, using the country mentioned in the category column as a way to filter news on certain countries has to be taken with caution. For example, a news about a pest in USA may mention previous research done in Spain, or the web may include 'Spain' in a list of previous posts (e.g. in a FAO web page). This seems to limit the possibility to infer spreading of PHT among countries unless other more precise geolocalization is available in the files.

Of the 148 useful news, only 22 were also obtained by the running -PHT categories. Even for the 16 news items that expressed an alert, only 5 were also obtained by a PHT category, thus further confirming the usefulness of the general new_pl_pests category to obtain news items on non-monitored PHT (Table 24).

Table 24: Analysis of the contents of useful news obtained in 2015 with the new_pl_pests category referring to the United Kingdom.

	general	***disease***	***name***	***alert***	**TOTAL**
useful news	53	22	57	16	148
also obtained by any -PHT category	3	1	13	5	22

Table 25 details the 16 news classified as 'alert' and that would point to some new PHT. However, from reading the MedIsys texts from 2015 and 2016 referring to 'United Kingdom', we could not identify a large increase in the number of news regarding to any of those 'alerts' after they were published (a maximum of 7 news items for the emerald ash borer in 2016).

Table 25: Text of 16 news that were obtained with the new_pl_pests category during 2015 and that expressed an alert for new PHT in the United Kingdom.

This *mysterious 'egg'* had some internet users baffled - after they believed it was an hatching in a field in Britain. 'alien life form' The weird jelly-like egg was found in the New Forest in the south of England by conservationist Dan Hoare, who posted pictures of his discovery on Twitter.
Trees in Tooting Common have become infested with a *highly contagious disease* which can cause branches to fall, disfigurement and bleeding bark. officers have identified that about 20 horse chestnuts lining Chestnut Avenue have been infected with bleeding canker disease, although there are fears all of the trees along the road could be infected.
Juniper, one of Scotland's most loved and treasured plants, is in *serious 'critical' decline and being killed off by a deadly new disease*, according to a new survey. Observations reported by the conservation group, Plantlife Scotland, suggest that 79 per cent of juniper in 2014 was mature, old or dead.
An aphicide has been granted an emergency authorisation to help oilseed rape growers *prevent the spread of a yield-sapping viral disease*. Teppeki (flonicamid), an aphicide from Belchim Crop Protection, was approved by Chemicals Regulation Directorate for the control of the peach potato aphid (Myzus persicae) in oilseed rape.
Thousands of Courier Country trees are being *checked for signs of a deadly disease*. An alert was raised after four cases of phytophthora ramorum were found in larch trees in Tayside. The disease attacks the wood, killing the trees. Three of the cases are in forests on private land near Forfar and Dundee, while the fourth is near Perth.
Phytophthora ramorum, whose first name means literally "plant destroyer", was first found in the UK at a garden centre in Sussex in 2002 and was first found in Wales just five years ago. It has not yet been found on trees in Scotland, but the fungus-like pathogen has been detected in the south-west of the country.
Sites affected by Hymenoscyphus fraxineus fungus include areas of woodland near Ambleside and Keswick, with disease already widespread across Europe. A tree infected with ash dieback disease. The *disease has spread widely across Europe* since trees were first reported dying in large numbers in Poland in 1992.
An invasive caterpillar, which munches through hedges and other plants, is crawling out of London and into the rest of the U.K. Experts are *concerned about the devastation it will cause*. The insect of concern is the box tree caterpillar, which is the larval stage of a moth native to the Far East and India.
An *invasive caterpillar* capable of reducing garden hedges to bare skeletons is *spreading from London across the UK*, experts warn. The gluttonous Asian box tree can gnaw its way through box hedges within days of hatching, wreaking havoc on prized gardens.
Parasitic "hitchhiking" moths, which infect and destroy the leaves of horse chestnut trees, are *moving north and could soon invade*. The horse chestnut leaf-mining moth, which originates in the Balkans, was first recorded in London in 2002 and *has spread throughout England and Wales*.
Potato growers are being urged to be *alert against an increasing late blight threat* in their crops as conditions turn to favour the moisture-loving disease. Predominantly dry weather has kept a lid on the devastating disease so far this season, allowing growers to keep their crops blight-free.
Chestnut trees are *under threat after an outbreak* of Asian wasps was *spotted for the first time in the UK*. The Forestry Commission issued an alert yesterday after a sighting of the oriental chestnut gall wasp (OCGW) was confirmed at Farningham Woods, Kent.
Britain's hedges under threat as South American caterpillar spreads box blight fungal disease. Daily Mirror Monday 27th April, 2015. Gardening experts have warned that *Britain's traditional hedges are under threat* - from a voracious South American caterpillar.
England's wine industry under threat from *four devastating plant viruses found in the country for the first time* RHS identify four separate viruses which can obliterate grape crops Experts say the only way to deal with problem is to pull up plants; Disease spotted on vines at RHS gardens in Wisley. ,....
An invasive beetle that has destroyed tens of millions of ash trees in the US *could pose a lethal threat to struggling native trees in the UK*. The emerald ash borer which arrived in Moscow seven years ago presents a serious threat to ash trees in Europe, researchers have warned.
Fruit growers are being warned of the *potential invasion of a pest* that has devastated crops in America and is now moving closer to the UK from southern Europe. The non-native brown marmorated stink bug (Halyomorpha halys) "has yet to come to England, as far as we know, but it has a devastating....

3.2.3. Categories based on Symptom-Expressions

This is an alternative approach to generate MedISys categories that do not contain threat names and are thus capable of detecting unknown threats. It is based on terms associated with 7 of the most

active threats, for which a rich model based on the ontology has been generated that includes symptoms, affected plant parts, vectors and affected crops.

The symptom expression part of the ontology has also been enriched since IR3 with additional translations for the identified symptoms. There are 37 symptoms extracted from the CABI forms as described in IR3. For IR3, the ontology featured 114 labels. Currently, at the end of the project, the ontology features 356 distinct labels for symptoms as listed detailed in Section 2.1.3.

There are translations for most of the languages and symptoms. In some cases, some of them have been removed because they were too ambiguous and generating too much noise when used as MedISys keywords, even when combined with host plants names and plant parts.

For instance, "dried", and the corresponding translations in other languages, have been removed because when combined with the plant part "fruit" they were selecting mostly news items unrelated with plant health threats. In all cases, we have kept the translations more specific to plant health threats, like "mummification".

The distribution of the translations among the considered languages is summarised in Figure 13.

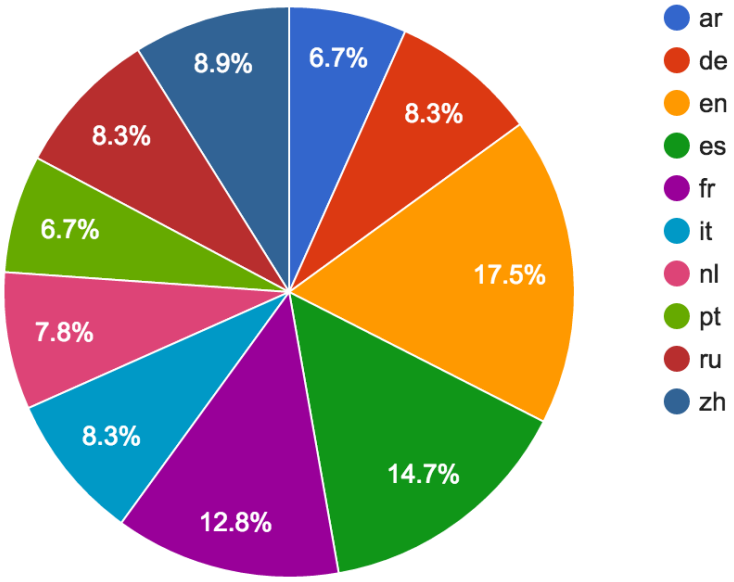


Figure 13: Percentage of symptom terms per language from a total of 356 terms (ar – Arabic, de – German, en – English, es – Spanish, fr – French, it – Italian, nl – Dutch, pt – Portuguese, ru – Russian and zh – Chinese)

The symptoms are combined with plant parts to model symptom expressions. This way, it is possible to generate MedISys word combinations that are more specific and less sensible to noise. The ontology now features translations for all the 10 languages under consideration for the 6 plant parts it models, as also detailed in Section 2.1.3. Overall, there are 96 terms for plant parts that are distributed among the languages under consideration as shown in Figure 14.

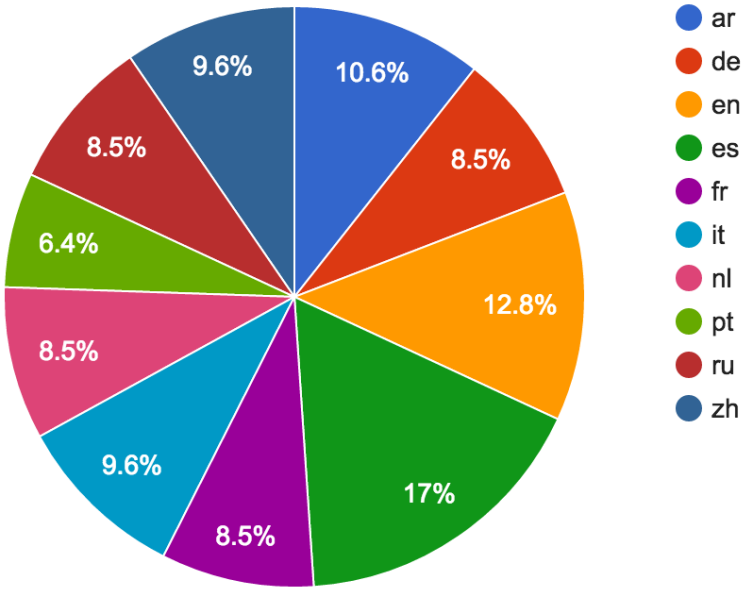


Figure 14: Percentage of terms related to plant parts for every language under consideration (ar – Arabic, de – German, en – English, es – Spanish, fr – French, it – Italian, nl – Dutch, pt – Portuguese, ru – Russian and zh – Chinese)

The previous sets of symptoms and plant parts were combined to model symptom expressions for the 7 pests that were completely modelled including affected crops, symptoms expressions and vectors. These pests are:

- *Phytophthora ramorum*, *Anoplophora glabripennis*, *Bactrocera tryoni*, *Agrilus planipennis*, *Xylella fastidiosa, Candidatus liberibacter* and *Rhynchophorus ferrugineus*.

For them, categories that do not include the name of the pest were generated. These include MedISys combination trees of affected crop plus symptom expression (i.e. symptom and affected plant part) and combinations of affected crop plus vector, as detailed in Table 26.

Table 26: MedISys combination trees to define the 7 categories based on symptom expressions

Combinations tree	**Affected crop** *AND* **Symptom** *AND* **Plant Part**
	OR
	Affected crop *AND* **Vector**

All these categories are included in the file accompanying this final report "ES3 - EFSAPlantHealthOntology-Symptoms.xlsx". The new version of this file includes many more combinations thanks to the previous enrichment of the ontology with more translations. From the original 221 combinations generated for IR3, the 7 categories are now based on 339 combinations.

These 7 categories are available from the accompanying file "ES3 - MedISysCategories-EFSAPlantHealthOntology-Symptoms" and its results are live at MedISys from the following links:

- http://medisys.newsbrief.eu/medisys/alertedition/en/AgrilusPlanipennis-PHT-Symptoms.html
- http://medisys.newsbrief.eu/medisys/alertedition/en/AnoplophoraGlabripennis-PHT-Symptoms.html
- http://medisys.newsbrief.eu/medisys/alertedition/en/BactroceraTryoni-PHT-Symptoms.html
- http://medisys.newsbrief.eu/medisys/alertedition/en/CandidatusLiberibacter-PHT-Symptoms.html
- http://medisys.newsbrief.eu/medisys/alertedition/en/PhytophthoraRamorum-PHT-Symptoms.html
- http://medisys.newsbrief.eu/medisys/alertedition/en/RhynchophorusFerrugineus-PHT-Symptoms.html
- http://medisys.newsbrief.eu/medisys/alertedition/en/XylellaFastidiosa-PHT-Symptoms.html

In order to analyse the potential of symptoms to identify new or unknown PHT we also manually created categories with the symptoms of Xylella in Italian as we expected to have many news on it. The two categories were:

sym_oak_xylella_it: a combination of the major affected crops or host plants together with '*Xylella*'.

combination			comment
	proximity	25	
	or	xylella	
	or	aceituno oliva olive oliveti olivi olivicoltori olivicultori olivo uliveti ulivi ulivo vite viti	list of known host crops

sym_oak_xylella_it_only: without mentioning *Xylella*, and using 3 combinations of symptoms. Symptoms were obtained from reading several web pages and documents obtained by a Google search written in Italian.

combination			comments
	proximity	25	
	or	aceituno oliva olive oliveti olivi olivicoltori olivicultori olivo uliveti ulivi ulivo	list of hosts
	or	abbassamento bancarotta battere caduta catastrofe celere colpire coniare crac crollo decadenza declino deperimento deterioramento deturpare devastazione disastro disgrazia dissesto distruggere fallimento fulmineo guastare iattura maceria malora mortalita mortalità perdizione picchiare repentino ribasso rovina rudere sciupare scoraggiamento sfascio tracollo tragedia vestigia avversità epidemia infetta infette infezione malattia morire morte parassiti patologia	key-words related to loss
combination			
	proximity	25	
	or	acacia aceituno acero agrumi alaterno ciliego citrus ginestra magnolia mandorlo mirto oleandro oliva olive oliveti olivi olivicoltori olivicultori olivo olmo pesco platano polygala prugno rosmarino spartium uliveti ulivi ulivo vinca vite viti	enlarged list of hosts
	or	ciacalina cicala cicale cicalella cicalelle cicaline oncometopia oncometopie rodrilegno saliva schiuma spuma spumosa sputacchina sputacchine sputo	popular names for the vectors
combination			
	proximity	25	
	or	acacia aceituno acero agrumi alaterno ciliego citrus ginestra magnolia mandorlo mirto oleandro oliva olive oliveti olivi olivicoltori olivicultori olivo olmo pesco platano polygala prugno rosmarino spartium uliveti ulivi ulivo vinca vite viti	enlarged list of hosts
	or	abbruciacchiato accrescimento+ridotto bruciati bruciatura bruscatura bruscature bruscitura clorosi+variegata declino deperimento dissecamenti dissecamento giallo imbrunimenti imbrunimento imbrunire ingiallimento ridotto+accrescimento	symtoms of injury or damage

	or	alberi albero branca branche chioma foglia fogliari foglie foglio foliare fusta fusti germogli lamina lámina legno lembo pianta piante rameali rametti rami ramificazioni ramo tronco vascolare	plant parts

In this category, the crossing of symptoms and plant parts was very broad, and not as detailed as obtained from the ontology.

Both definitions are available at:

- http://medisys.newsbrief.eu/ medisys/alertedition/es/sym_oak_xylella_it.html
- http://medisys.newsbrief.eu/ medisys/alertedition/es/ sym_oak_xylella_it_only.html

Evaluation

The evaluations carried out show that the categories based on symptoms expressions are very susceptible to noise because they do not use the pest or threat name and rely on words related to symptom expressions, such as plant parts or symptoms. These terms come from the ontology as previously detailed and unfortunately have very different meanings, in most cases not related to plant health threats. For instance, "leaf" or "death".

Consequently, to avoid too much noise, the experiments focused on reducing the "Proximity" parameter of these categories, so if we are looking for the "dead leaves" symptom expression, we look for combinations of words like "death" or "dead" plus "leaf" or "leaves" that appear next to each other. This largely reduced the noise while not significantly reducing the recall because these terms lose their meaning as a symptom expression if they appear separate in the news item.

Additionally, the evaluations allowed collecting a set of negative words, terms that if they appear in a news item then it is not selected by the category, that avoid typical cases we have encountered related to pharmacy or cooking sites, in many cases including hacked sites or those including unrelated keywords indiscriminately to attract traffic.

For example, the list of negative words for English are:

pill, pills, viagra, pharmacy, recipe, recipes, cook%, war, troop%, militar%

And for Spanish:

p_ldora, viagra, farmacia, receta, recetas, cocina%, guerra, militar%

The combination of these two features, negative words and a small "Proximity" of 15 words or less, reduced the noise and also the amount of news items captured by these symptom-based categories. Currently, they are producing on average 1 news item each per week and approximately 60% of them are potentially relevant, as described in the rest of this subsection, which shows and examples of this kind of results. However, this low volume makes it potentially manageable to deal with the amount of noise that hardly is going be possible to reduce due to the ambiguity of the terms used with a keyword based search engine like MedISys.

In fact, the volume is so low that the category for *Phytophthora ramorum* symptoms just generated one hit during the period analysed. However, it was a very interesting one because it was related with a different pest which shares some of the *P. ramorum* symptoms, *Phytophthora hydropathica*. Moreover, the source is the New Disease Reports journal so clearly a good candidate for a new plant health threat.

First report of Phytophthora hydropathica in river water associated with riparian alder in Spain
ndrs Friday, June 10, 2016 6:49:00 PM CEST | info [other]

Trigger words: [PhytophthoraRamorum-PHT-Symptoms] lesions[1]; wilting[1]; dieback[1]; bark[1]; necrosis[2]; Rhododendron[1]; leaf[1]; stem[2]; leaves[6];

Other categories: Sirococcus tsugae;

Phytophthora hydropathica has been commonly reported from riparian sites in southeastern USA, on watersheds and nursery sites in Tennessee (Hulvey et al ., 2010) and in nursery irrigation reservoirs in Virginia (Hong et al ., 2010). Recently it was also recovered from soil associated with Viburnum tinus in Italy (Vitale et al ....

Next, additional examples of results of these categories analysed during the evaluation are presented. Some of them are news items that were not selected by other categories related to plant health threats. Consequently, these are the results more prone to irrelevant results but also those more potentially valuable because they would be ignored otherwise.

The next results shows a clear example of a relevant news item, in this case from EPPO, related to a particular pest that is selected by the combination of symptoms expression for a different one because there is some overlapping.

Bactrocera latifrons
eppo Thursday, November 19, 2015 3:14:00 PM CET [other]

Trigger words: [BactroceraTryoni-PHT-Symptoms] Citrus[1]; fruit[10]; rot[1]; tomato[3]; fruits[4]; Tomato[1]; Fruit[1]; Passiflora[1];

Where: B. latifrons originates from Asia but its range has expanded through introductions into Africa (Kenya and Tanzania, first found in 2007 and 2006 respectively) and the islands of Hawaii (US, first found in Honolulu in 1983) and Yonaguni (Okinawa prefecture, Ryukyu Archipelago, Japan, first found in 1984)....

Next, there is an example of an also relevant news item selected by both, the filter based on the threat name and the one based on the symptoms:

El sector del aceite de oliva, en alerta por la expansión de la bacteria Xylella
abc-Andalucia Monday, December 14, 2015 3:38:00 AM CET | info [en] [other]

Trigger words: [XylellaFastidiosa-PHT-Symptoms] oliva[2]; olivo[4]; olivar[1]; árbol[1]; Oliva[3]; muerte[1]; hojas[2];

Entities: Ramón Díaz[1];

Other categories: Xylella fastidiosa;

por las gravísimas consecuencias que la bacteria Xylella fastidiosa puede acarrear en unos de los buques insignia de nuestra agricultura. Los estragos que ha causado en Italia, concretamente en la región de Apulia, al sureste del país, y que obligan a arrancar el olivo como una solución eficaz para.....

In addition to clearly relevant items, it was possible to also spot items that might not seem relevant at a first glance but that might become potentially relevant in some particular context, like tryinf to track the spread of a pest or disease. For instance, the AgrilusPlanipennis-PHT-Symptoms category selected the following news item, which is not selected by any other MedISys category.

Area update: St. Joe haunted trail benefit set for Saturday

news-gazette Friday, October 30, 2015 3:37:00 PM CET | info [other]

Trigger words: [AgrilusPlanipennis-PHT-Symptoms] Ash trees[1]; ash trees[1]; Tree[1]; tree[1]; dead[1];

ST. JOSEPH — For years, Mary Derenne and her family transformed their yard into a haunted trail and asked visitors to donate to St. Jude Children's Hospital. This year, the location, duration and beneficiary have all changed. The event, which will take place in Crestwood Park in St. Joseph, is for one night only: 6 p....

This might not seem relevant but, actually, this result points to a web site that includes many news items, all related to local news but with very different topics. The news item that really triggered the AgrilusPlanipennis-PHT-Symptoms category is shown next, highlighting the triggering words:

GIBSON CITY COUNCIL

Ash trees to be removed

GIBSON CITY — City Superintendent Randy Stauffer said there are five ash trees that are hazardous because of dead top branches and hollow centers. Council members approved spending $11,500 to have Michael Poor, a certified arborist from Urbana, remove them.

Stauffer felt the price was fair, saying similar past bids have run as much as $3,000 per tree.

Poor will work with Tom Barrow of Barrow Tree Service in Gibson City to haul away the debris at no charge.

Poor's fee includes everything except stump removal. Stauffer said Barrow could be hired for that work at $97 per stump.

JEAN NOELLSCH

Paxton Record

As it can be observed, this news item might not be directly reporting about a plant health threat but might be interesting to take it into account if conducting research about reported ash trees that might be potentially related to the pest.

Finally, given the high level of noise of these categories, there were many clearly irrelevant results, not related to plant health threats. However, it is reported to show the difficulties arising from using just symptom expressions and not threat names. However, results like this one have shown to be useful as a source of negative words to reduce noise.

Israel: Palestinian Assailant Killed, 2 Protesters Dead

ABCnews Friday, December 11, 2015 4:46:00 PM CET | info [other]

Trigger words: [XylellaFastidiosa-PHT-Symptoms] olive tree[1]; tree[1]; death[1]; stem[1]; olive[1];http://emm.newsexplorer.eu/NewsExplorer/entities/en/459415.htm

A Palestinian man plants an olive tree during demonstration on the anniversary of of the death of Palestinian cabinet minister Ziad Abu Ain, who collapsed shortly after a protest on Dec. 10, 2014, in the West Bank village

of Turmus Aya, as demonstrators clash with the troops near the village outside of Ramallah, Friday, Dec....

The selection keywords are very unrelated to the news item because they are really part of the caption accompanying the article, as shown next.

Israel: Palestinian Assailant Killed, 2 Protesters Dead

By DANIELLA CHESLOW, ASSOCIATED PRESS

JERUSALEM — Dec 11, 2015, 11:59 AM ET

A Palestinian man plants an olive tree during demonstration on the anniversary of of the death of Palestinian cabinet minister Ziad Abu Ain, who collapsed shortly after a protest on Dec. 10, 2014, in the West Bank village of Turmus Aya, as demonstrators clash with the troops near the village outside of Ramallah, Friday, Dec. 11, 2015. (AP Photo/Majdi Mohammed)

The brother of a Palestinian teen who died in unclear circumstances in October was among three Palestinians — including one suspected attacker — killed by...
...
Israel blames incitement by political and religious leaders for the violence. Palestinians say the attacks stem from despair at achieving statehood.

This wrongly selected news item was used to generate additional negative words like "war", "troop%" or "militar%".

Finally, it was also possible to identify cases showing the usefulness of monitoring, in addition to symptoms expression, the plant health threat vector. In this case, though the news item was selected by other categories because it also includes "Huanglongbing", it is interesting to note that it was selected by the corresponding symptoms category because this kind of filter also includes the vectors associated to the plant health threat, in addition to the affected crop.

FMC launches Mustang 350 EG for citrus in Brazil Dec. 11, 2015

agropages Thursday, December 10, 2015 5:29:00 PM CET | info [other]

Trigger words: [CandidatusLiberibacter-PHT-Symptoms] Citrus[3]; citrus[7]; Diaphorina citri[1]; Asian Citrus psyllid[1]; smaller[1];

Other categories: New Plant Pests; Insecticides; Multiple Species;

FMC Agricultural Solutions announced this Monday (12.07) the launch of the insecticide Mustang 350 EC for citrus. The focus of the new product is to control the moth Gymnandrosoma aurantianum and the Asian Citrus psyllid (Diaphorina citri) two of the main plagues that affect citrus in Brazil....

Finally, as for the 2 manually created categories for Xylella symptoms in Italian, results were analysed at two sample dates, October 2015 and June 2016. During October 2015 (24th to 28th), the symptoms only category (sym_oak_xylella_it_only) identified 60 news: 20 (30%) were related to

Xylella and were also triggered by both categories mentioning *Xylella* (sym_oak_xylella_it and XylellaFastidios-PHT).

For the 2016 sample (April to June) the category with only symptoms (sym_oak_xylella_it_only) identified 104 news: 61 (61%) were related to *Xylella*, and they were also triggered by both categories because the name, Xylella, was also present in the text. But the symptoms category did not search for the name, and therefore it showed its potential when identifying theats using only symptoms and not specific names.

Moreover, many news identified by the manually curated symptoms category (sym_oak_xylella_it) were not triggered by the existing category for *Xylella fastidiosa* (XylellaFastidiosa-PHT). This was because many newspapers popularised the name 'Xylella' without using the full scientific name. This has now been corrected. This is something to be aware of when creating new specific pests categories: they should also include the generic name. Although this can result in greater noise, it can be expected that newspapers will not publish many news about related species if they are not plant health threats.

3.3. Evaluation of MedISys Monitoring Reporting

To conclude the evaluation of the results of the project, the focus was placed on the current approaches and strategies for reporting the identified signals to the EFSA Units and experts through the MedISys interface. This study is based on common practices in the User Experience evaluation community (Nielsen, 1994). The evaluation is based on a set of representative user tasks available through the MedISys user interface based on the experience of plant health experts. A set of representative users were then asked to perform the previous set of user tasks. Their interaction was recorded and then analysed using User Experience evaluation techniques. The whole process and the outcomes are reported in this section.

It is also important to note that the MedISys user interface allows users to register their interest about particular categories so they are informed by e-mail in case of new items are captured by the corresponding category. The evaluation also tests the registration process, while the usefulness of the e-mail alerts has been evaluated using a survey included in Annex A. In this case, 29 users from EFSA staff and PLH panel were subscribed to a combined alert for all 47 named threat categories in Table 1. From December 2015, they received the daily email alerts, with all news items captured by any of the 47 named threat categories. After more than 6 months of subscription, they were asked to respond to a survey. Almost half of them, 14 users, responded to the survey and all of them reported that they continue to be interested and are still registered. Half of these 14 users consider the alerts useful for their daily work, while the other half do not. The main concern is that a daily alert is too frequent and they would prefer a weekly one, preferably curated (the current one is automatically prepared, thus inevitably including noise and redundant items). Additional details about the survey are reported in Annex A.

3.3.1. User Tasks for User Experience Evaluation

Fourteen user tasks were defined. They are based on what is available from the MedISys user interface for end-users and the experience gained by plant health experts while using it. The tasks are divided in two categories. From Task 1 to Task 3, these are the "Generic Tasks" carried out at the MedISys EFSA page level. Then, from Task 4 to Task 14, they are the "Specific Tasks" related specifically with pests and performed at the level of the pages related to pests involving insects.

Generic Tasks

These tasks are carried out at the level of the entry page for EFSA available from MedISys. The link to this page is:

http://medisys.newsbrief.eu/medisys/categoryedition/symptoms/en/efsa.html

Task 1. What are the latest news related to EFSA?

Task 2. What and where are the most active topics about plant health threats?

Task 3. Regarding Spain, which have been the latest news related to EFSA?

Specific Tasks

The rest of the tasks consider a particular set of alerts in the EFSA section of MedISys, concretely those related with plant health threats associated to insects. The entry page for this kind of pests is:

http://medisys.newsbrief.eu/medisys/groupedition/en/PlantHealthInsects.html

Task 4. What have been the latest news for pests related with insects?

Task 5. Please, create an alert to receive notifications related to this type of pest by e-mail. Please, do the same but to get them through a news feed based on RSS.

Task 6.Which have been the most active pests of this kind?

Task 7. Which are the countries with more activity of this kind? And for the most active pest?

Task 8. Please, narrow the selected news from those worldwide to just Europe, in the Mediterranean countries and in the EPPO countries.

Task 9. Browse all the news in MedISys for the last 24 hours in a particular country, for instance Spain.

Task 10. Create an alert to receive notifications about news items about plant health threats in the previous country.

Task 11. Select a particular news of your interest and try to answer the following questions: In which country does it take place? Is this country the same where the news item was written?

Task 12. Please, try to display a particular news on a map. Now, try to do the same with a set of news.

Task 13. What are the potential new pests identified by MedISys? Have there been warnings of this kind in Spain?

For a list of news about potential new pests you can use the link:

http://medisys.newsbrief.eu/medisys/alertedition/en/oak_new_plant_pests.html

Task 14. Create an alert to be notified if there are such alerts in the future? Create another one that focuses on Spain.

3.3.2. Testing Equipment and Involved Users

The user experience evaluation was performed on March 15th 2016 with the participation of 3 experts involved in the project and on June 20th 2016 with 2 experts from the EFSA PLH Panel. Based on existing practice, 5 experts are enough for an evaluation that focuses just on effectiveness and does not require and efficiency study (Nielsen, 1995).

For the first evaluation session, the usability evaluation was done at the UsabiliLAB, the usability laboratory that the GRIHO research group has at the Universitat de Lleida. The UsabiliLAB is a specific room to develop R+D projects and technology transfer projects in the User Experience evaluation domain. The second session was performed at the EFSA building in Parma.

The UsabiliLAB equipment used during the first evaluation was:

- Eye Tracker: Eye Tracking is the process of measuring either the point of gaze (where one is looking) or the motion of an eye relative to the head. An Eye Tracker is a device for measuring eye positions and eye movement.
- Morae: specific software to carry out and record a User Experience evaluation session. This tool also facilitates analysing the recording to measure the effectiveness and efficiency of the user while performing the proposed user tasks.

For the second evaluation, the Eye Tracker was not available. Once the previous equipment was set, and for each participant, they were asked to perform all the user tasks listed in the previous

subsection under the supervision of an evaluator. Figure 15 and Figure 16 show some moments and screenshots of the assessment.

Figure 15: Overview of the gaze activity of one of the participants during one of the tasks. The left image is a heat map, where hot areas correspond to those that attracted more user attention. The right image is a gaze plot that shows the path followed by user gaze.

Figure 16: Detailed view of the top part of the images in Figure 15. Left image: heat map. Right image: gaze plot.

3.3.3. Evaluation Results

The evaluation focused on User Experience effectiveness. Therefore, what was measured for each user was the completion level for every task, ranging from 0%, if no part of the task was completed, to 100%, if the task was completed entirely.

Table 27 presents all effectiveness measures for all users and tasks together with comments noted by the evaluator during the evaluation session or later, while reviewing the session recordings.

Table 27: User Experience effectiveness results for each user and task, ranging from 0%, if the user was unable to complete any aspect of the task, to 100%, if it was completed entirely.

	USER 1		USER 2		USER 3		USER 4		USER 5	
TASK 1	50%	Reached EFSA page but unable to locate the related news	100%	Completed but not efficiently	100%		100%	Hard to find menu option, hidden at the bottom	50%	Unable to find menu option, hidden at the bottom
TASK 2	50%	Not able to identify all most active topics, confused by colours	100%		0%		50%	Looking for specific page for PHT. Finally went to the general one.	0%	Missing specific page for PHT aggregating all PHT categories
TASK 3	50%	User abandons finding task too complicated	0%		0%	User tries for a while but finally abandons	100%	Got map is clickable	100%	Used map to click
TASK 4	100%	The task is evident for this user and completed efficiently	0%	The user tries to use the search form but finally abandons	100%		100%		100%	
TASK 5	50%	Completed e-mail alert but unable get RSS, found complicated for non tech people	50%	Completed e-mail alert but unable get RSS, found complicated for non tech people	70%	Almost the whole task was completed, but some concepts were confusing	50%	OK, but not RSS	50%	OK, but not RSS
TASK 6	30%	Reached results but unable to interpret them. Some news are older and the graph is hard to understand	20%	Unable to locate the user interface component to define the filter, too small and unintuitive	0%	Good start, but after trying to filter and sort the user abandons, does not know what more to do	0%	Didn't find aggregate view for this kind	0%	Didn't find aggregate view for this kind
TASK 7	40%	Confused by the alert statistics graph	70%	Task completed almost in its entirety after starting from a specific insect	0%	Good start, but after trying to filter and sort the user abandons, does not know what more to do	0%	Missing map for aggregates like insects. More than 1 month old not appearing in map.	0%	Missing map for aggregates like insects
TASK 8	0%	User tries to use the map but it does not help to perform this task	0%	User tries to use the map but it does not help to perform this task	0%	User tries to use the map but it does not help to perform this task	33%	OK specific country. No way to focus on Mediterranean or EPPO countries.	33%	OK specific country. No way to focus on Mediterranean or EPPO countries.
TASK 9	40%	The user gets frustrated half way because the user finds it harder than expected	100%		0%		0%	The top menu name "Continents" is not indicative	100%	Used continents
TASK 10	20%	Font size is too small and difficult to read. After filtering by country, the user was unable to complete the task.	100%		0%		0%	Tried to click active topics per country but not clickable	50%	No way to filter topics. Considered better to go reverse using the Country but confusing "from" "about". Too small
TASK 11	30%	The user got lost, not capable to identify where in the interface she/he is and abandons	50%	The user is capable of identify news and country but unable to identify their origin	30%	The user got lost, not capable to identify where in the interface she/he is and abandons	50%	Flags as indicative of where published. No way to see countries mentioned.	0%	Missing way to see the countries mentioned by a news item
TASK 12	0%	The map does not work and KML is unavailable.	0%	The map does not work and KML is unavailable.	0%	The map does not work and KML is unavailable.	0%	Click MAP link, but not working	0%	Not working

TASK 13	0%	The user was unable to guess how to complete this task and abandoned	0%	The user was unable to guess how to complete this task and abandoned	30%	The user is able to complete part of the task but soon find it too complicated and abandons	0%	Map not being shown	33%	OK
TASK 14	20%	The user is able to start the task filtering by pest but is unable to guess how to continue	0%		0%		50%		50%	
AVERAGE	**34%**		**42%**		**24%**		**38%**		**40%**	

As can be observed from the effectiveness results reported in the previous table, the effectiveness measures were very low for all users. The average for all users was 36%. Usually, a user interface would be expected to show 78% effectiveness to avoid user frustration (Sauro, 2011). Consequently, in general, it can be said that the MedISys website user interface was hard to use for end-users that have not received specific training. The number of unfinished tasks highlights this aspect. Even considering that only five users participated in the evaluation, enough taking into account that this is just an effectiveness evaluation involving experts (Nielsen, 1994), most of the tasks were impossible or too hard to complete.

Even after investing a lot of effort, it was not evident how many of the proposed tasks can be completed using the MedISys user interface. These tasks were identified by the involved experts as very relevant from a plant health threats monitoring perspective.

Next, we propose what might improve the User Experience. First of all, a more uniform user interface should be presented. Currently, just categories for individual threats (for instance *Agrilus planipennis* in Insects) have all the user interface widgets activated (RSS, Map, Alert Levels, Most reported countries, Daily number of articles in this category, Most active sources...). In this case it is possible to use the map to click countries and enable geolocation at the country level to monitor, for instance, pests spread across countries through its impact in the media, as shown in the left screen capture in Figure 17. However, the right screen capture also in Figure 17, for a different pest, does not provide the map view and thus geolocation-based filtering is not possible in this case.

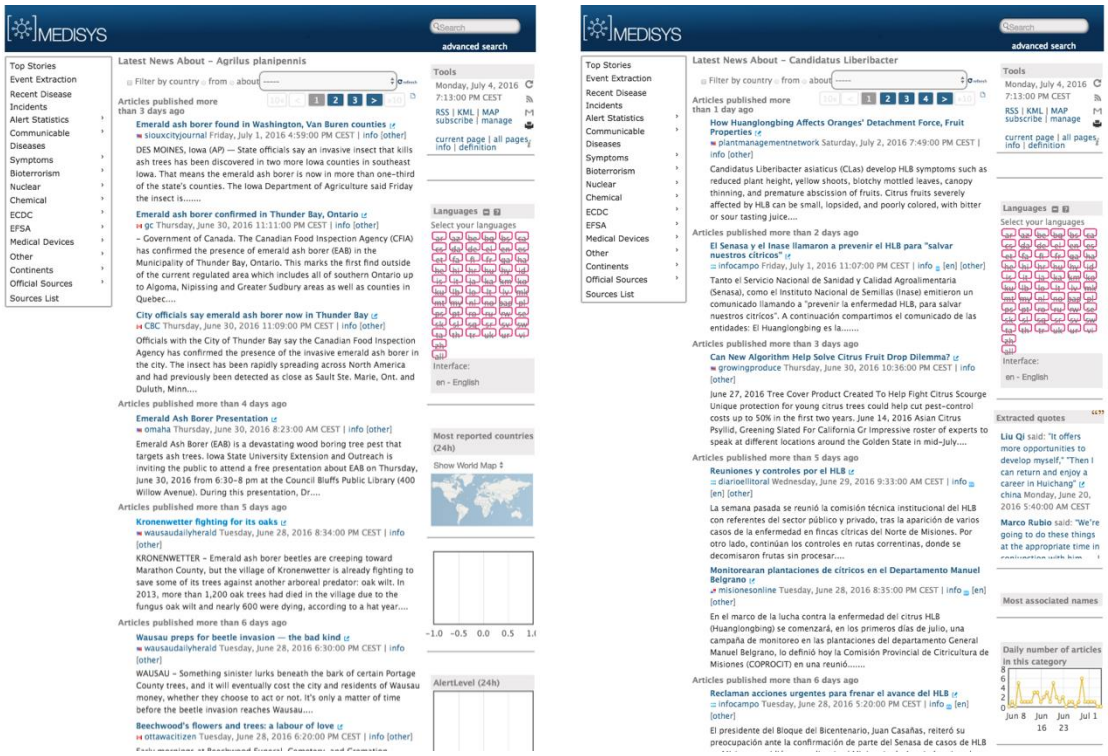


Figure 17: Comparing two screen captures of MedISys reports for two different pest categories generated during the project. Left, for *Agrilus planipennis*, displays most MedISys visualization widgets including map view. Right, for *Candidatus Liberibacter*, displays many visualization widgets but lacks the map view so geolocation filtering is not available.

This confuses users and makes it impossible for them to use these functionalities if they are interested in threats at a more generic level than a specific pest. The proposal is to enable these views at the aggregated level (for instance RSS or daily numbers of articles at the Insects level). Moreover, it would be also helpful to have an overview of the situation at the plant health threats level. This can be attained by enabling these widgets at the global view called "PlantHealth" that summarises the results for all the categories related with plant health threats.

The idea would be then to have 3 levels:

- Level 1. Plant Health (top level), with "Most Active Topics" for al Plant Health categories, just like it is available at the more generic EFSA level.
- Level 2. Groups based on pest type, 7 submenus (PlantHealthBacteria, PlantHealthFungi, PlantHealthInsects, PlantHealthMollusks, PlantHealthNematodes, PlantHealthOomycetes and PlantHealthVirus)
- Level 3. Individual pages for each plant health threat category.

For each level, the user interface would present the aggregated view of all the levels below through the same user interface components that are available for the individual threats (i.e. RSS, Maps, plots...). In order to further personalise the aggregated levels and define which sublevels should be considered, an interesting feature would be to also add a facet to select the sublevels to be considered. For instance, if the user is in the PlantHealtInsects level 2 aggregation, it should offer the user the option of selecting all or just some of the individual plant health threats related with insects and show the aggregated view for them.

This way it would be also possible to help users to get a personalized overview of the most active categories at the whole plant health level. It is true that NewsDesk offers this functionality, but as this is not available for all users, it might be interesting to provide this aggregated view at the whole Plant Health level and also at the threat type level (bacteria, insect...). This way, it is also possible to make use of the map views that highlight the most active areas for plant threats in general.

Other desirable features highlighted by the involved users were:

- Prioritize the news items based on the confidence in the source. Measures to derive this confidence are available from the source listings provided in previous deliverables.
- When showing selected news, in addition to the title and part of the text, display the text around the keywords that have been used to select the news item, just as search engines do.
- Improve the automated translations of the news items, especially for languages like Chinese or Arabic.
- Besides the existing maps for Europe, Africa and the Middle East, create a map with all EPPO countries in order to better visualize presence and spread of news within EPPO and neighbouring countries.

4. Conclusions and Future Work

From the work detailed in this report, and also described in a previous publication (Alomar et al., 2015), at the end of the project, all the objectives have been addressed. First of all, for Objective 1, a significant collection of news sources has been selected after an evaluation process that guarantees their timeliness so they can be monitored by MedISys. Two different approaches were followed to build this collection, a direct approach based on the manual selection of sources and an indirect one that uses Web search engines to identify them.

In the case of the manually curated information sources, their analysis has shown that they are highly relevant in terms of metadata and content quality. However, the amount of sources selected for monitoring using this method, 61, is relatively low. And, what is more important, it is going to be difficult to extend it because it is based on a manual process and already established knowledge. Therefore, these sources might not be the best ones to detect re-emerging and especially new emerging plant health threats.

On the other hand, with an indirect information sources collection method based on Web search, it has been possible to identify other previously unknown sources to be monitored, concretely 1884 information sources. Even at this early stage, the Plant Health Threat Ontology built to collect the knowledge about plant health threats during the project already proved its usefulness as the source of keywords to use for automated Web searches.

The information sources selected using both methods seem quite complementary and produce an interesting set of sources to be monitored by MedISys. Interesting in the sense that it combines well-known and high-quality sources that can serve as reference with unknown and less-quality information sources where it is more likely that re-emerging and new emerging plant health threats are detected.

In relation with Objective 2, the enrichment of the Plant Health Threat Ontology with pest and disease names coming from multilingual sources such as UniProt Taxon, EPPO or Wikipedia has allowed us to generate MedISys categories that monitor known plant health threats in the media. From an initial list of 140 candidates, 117 of them have been finally mapped to this multilingual sources. Consequently, it has been possible to generate 117 MedISys categories for known plant health threats based on weighted words lists including scientific names, other scientific names and common names in different languages, overall 1609 labels at the end of the project.

During the project, these categories have already proven to be very useful, providing mostly relevant results because they search for names of known pest so the selected news items are very likely relevant. This also supports the use of the ontology as a mechanism to organise the multilingual keywords associated to a plant health threat. Our recommendation is to follow this approach when adding additional categories to MedISys in the future. It is not mandatory to use an ontology management tool for that. It is enough to keep in mind the ontology structure presented in Figure 1 and Figure 2 when arranging the keywords using the MedIsys category editor.

The list of plant health threats under consideration for immediate inclusion as new categories in MedISys, beyond the end of the project, are:

- *Atropellis* spp (fungus)
- *Ceratocystis platani* (fungus)
- *Cryphonectria parasitica* (fungus)
- *Diaporthe* spp (fungus)
- *Ditylenchus destructor* (nematode)
- *Drosophila suzukii* (insect)
- *Eotetranychus lewisi* (mite)

- *Erwinia amylovora* (bacterium)
- Flavescence dorée (bacterium)
- *Ophiostoma novo-ulmi* (fungus)
- *Phyllosticta citricarpa* (fungus)
- *Phytophthora infestans* (oomycete)
- *Radopholus similis* (nematode)

On the other hand, for Objective 3, the approach has been to generate MedISys categories that are not based on threat names so they are able to detect news items mentioning unknown or unnamed threats. Two complementary approaches were followed in this case. The first one was to build a category based on a combination of words manually curated by an expert based on words typically present in document talking about new plant health threats. The other was based on terms associated with a selection of 7 of the most active threats that have been modelled with high detail in the Plant Health Threat Ontology. These terms are associated with the threat, but do not include any of its names. They are the associated vectors, the affected crops or the symptoms of the threat.

Two categories based on the first approach, manually curated terms, have been generated and tested. The results are quite satisfactory in this case, with about 80% of the selected news items relevant from a plant health perspective. The volume of selected news is quite high, about 10 per day. Consequently, a second category that also requires the presence of words associated to emergency situations has been generated, whose volume is largely reduced to about 1 per week.

An alternative approach to new-emerging threats category generation has been also explored. It is based on combining, for a particular pest or disease, the terms associated to the affected crop, the involved vectors and symptom expressions, which include symptoms and plant parts.

Seven plant health threats have been modelled with great detail in the ontology: *Phytophthora ramorum*, *Anoplophora glabripennis*, *Bactrocera tryoni*, *Agrilus planipennis*, *Xylella fastidiosa, Candidatus liberibacter* and *Rhynchophorus ferrugineus*. This has allowed us to generate categories that just feature terms associated with symptoms expressions, vectors and crops, without including the plant health threat names (scientific, common, etc.).

These categories have been also tested but due to the great level of ambiguity of the terms associated especially to symptoms expressions, the amount of noise is significantly higher. Towards the end of the project, it has been possible to identify some symptom expressions that were responsible for a big part of the noise, like "dried" combined with "fruit". They have been removed, keeping more specific terms like "mummification" and the relevance of the results has raised from about 50% to about 60%. The volume is about 1 per week per category.

It is important to note that these categories are highly affected by ambiguities in language because the terms used for symptoms expressions have a wide range of meanings, most of them not related to plant health. For instance, "trunk" or "reddening". Unfortunately, MedISys is a search engine based on keywords and not word meanings so it seems not possible to go beyond this relevance ratio. In any case, the work carried out might be an interesting starting point for future studies about how symptoms are expressed across pests and diseases or affected crops in order to generate a detailed network of plant health threats, like the one proposed for human health by Zhou et al. (2014).

To conclude, in relation with Objective 4, the study with plant health experts based on 14 of their typical information needs, it has been observed that although MedISys provides all the user interface components to fulfil them, just an average of 36% of the tasks were completed by experts.

For future work, and from what has been observed during the project, the proposal is to perform some improvements in the MedISys user interface. First of all, a more uniform user interface should be presented. Currently, just pages for individual categories (for instance AgrilusPlanipennis-PHT in

Insects) have all the user interface widgets activated (RSS, Map, Alert Levels, Most reported countries, Daily number of articles in this category, Most active sources,...).

This confuses users and makes it impossible for them to use these functionalities if they are interested in threats at a more generic level than a specific pest. The proposal is to enable these views at the aggregated level (for instance RSS or daily numbers of articles at the Insects level). Moreover, it would be also helpful to have an overview of the situation at the level of all plant health threats. This can be attained by enabling these widgets at the global view called "PlantHealth" that summarises the results for all the categories related with plant health threats.

For each level, the user interface will present the aggregated view of all the levels below through the same user interface components that are available for the individual threats (i.e. RSS, Maps, plots...). In order to further personalise the aggregated levels and define which sublevels should be considered, an interesting feature would be to also add a facet to select the sublevels to be considered. For instance, if the user is at the PlantHealtInsects level of aggregation, to offer the user the option of selecting all or just some of the individual plant health threats related with insects and show the aggregated view for them.

Finally, it should be also considered for future work to regularly update the list of sources monitored by MedISys, e.g. by running searches for new plant health threats using different search engines. The list of most frequent sources should then be checked against the list of currently monitored sources to identify the missing ones.

References

Alomar O, Batlle A, Brunetti JM, García R, Gil R, Granollers A, Jiménez S, Laviña A, Linge JP, Pautasso M, Reverté C et al., 2015. Development and testing of the media monitoring tool MedISys for early identification and reporting of existing and emerging plant health threats. *EPPO Bulletin*, **45**(2), 288-293.

Arsevska, Roche M, Hendrikx P, Chavernac D, Falala S, Lancelot R, Dufour B, 2016. Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Computers and Electronics in Agriculture,* **123**, 104-115.

EFSA, 2012. Evaluation of the MedISys EFSA Tab menu in the areas of animal and plant health, GMOs, pesticides and biological hazards. Supporting Publications 2012: EN-331. http://www.efsa.europa.eu/en/search/doc/331e.pdf

Lee YW, Strong DM, Kahn BK and Wang RY, 2002. AIMQ: a methodology for information quality assessment. *Information & Management*, **40**, 133-146.

Margaritopoulos T, Margaritopoulos M, Mavridis I and Manitsaris A, 2008. A conceptual framework for metadata quality assessment. Proceedings of the International Conference on Dublin Core and Metadata Applications, pp. 104-113.

Naumann F, 2002. Quality-driven query answering for integrated information systems. Lecture Notes in Computer Science, Vol. 2261. Springer, Berlin.

Nielsen J, 1994. Usability inspection methods. In: Proceeding of the CHI '94 Conference Companion on Human Factors in Computing Systems, ACM, pp. 413-414.

Nielsen J, 1995. How to conduct a heuristic evaluation. Nielsen Norman Group. Available online (last accessed November 2016) at: http://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation.

Salaún Y and Flores K, 2001. Information quality: meeting the needs of the consumer. *International Journal of Information Management*, **21**(1), 21-23.

Sauro J, 2011. What is a good task-completion rate? MeasuringU. Available online (last accessed November 2016) at: http://www.measuringu.com/blog/task-completion.php.

Stvilia B, Gasser L, Twidale MB and Smith LC, 2007. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology,* **58**(12), 1720-1733.

Zhou X, Menche J, Barabási AL and Sharma A, 2014. Human symptoms–disease network. *Nature Communications*, **5**, 4212.

Glossary and Abbreviations

DBPedia It is a project aiming to extract data from the Wikipedia and represent it using RDF.

Dublin Core It is a vocabulary of terms used to describe web resources and provide metadata about them.

Graph An RDF graph is a collection of RDF statements.

IQm Information Quality Method developed to rate the quality of online information.

Ontology An ontology formally represents knowledge within a domain using a concrete vocabulary to denote the concepts, properties and interrelationships among them.

RDF The Resource Description Framework (RDF) is a family of standards to describe metadata about web resources.

RDFa A standard to encode semantic data based on the RDF model into HTML content.

Reconciliation The process of integrating data from different providers.

RSS Stands for Rich Site Summary. It is a family of standard web formats to publish frequently updated content such as blog entries, news, audio, etc.

SPARQL Standard for query, retrieve and maniuplate data stored in RDF format.

Taxon A taxonomic group of any rank, such as a species, family, or class.

URI A Uniform Resource Identifier (URI) is a string that identifies a web resource.

Usability In software engineering, usability is the degree to which a software can be used by specified consumers to achieve quantified objectives with effectiveness, efficiency, and satisfaction in a quantified context of use.

User Experience Refers to a person's total experience using a particular product, system or service. It includes the practical, experiential, affective, meaningful and valuable aspects of human–computer interaction and product ownership. Additionally, it includes a person's perceptions of system aspects such as utility, ease of use and efficiency.

UX User Experience.

Virtuoso A database to store and retrieve triples, usually represented in RDF.

Appendix A – Source Metadata Properties (from Dublin Core Metadata Elements)

Parameter	Definition
Title	A name given to the resource.
Creator	An entity primarily responsible for making the resource.
Subject	The topic of the resource.
Description	An account of the resource.
Publisher	An entity responsible for making the resource available.
Contributor	An entity responsible for making contributions to the resource.
Date	A point or period of time associated with an event in the lifecycle of the resource.
Type	The nature or genre of the resource.
Format	The file format, physical medium, or dimensions of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Source	A resource from which the described resource is derived (i.e. URL/web address).
Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant. Includes Geographic coverage and Period coverage.
Language	A language of the resource.
Rights	Information about rights held in and over the resource. Y or N

Appendix B – Content Quality Attributes and Rating Scale

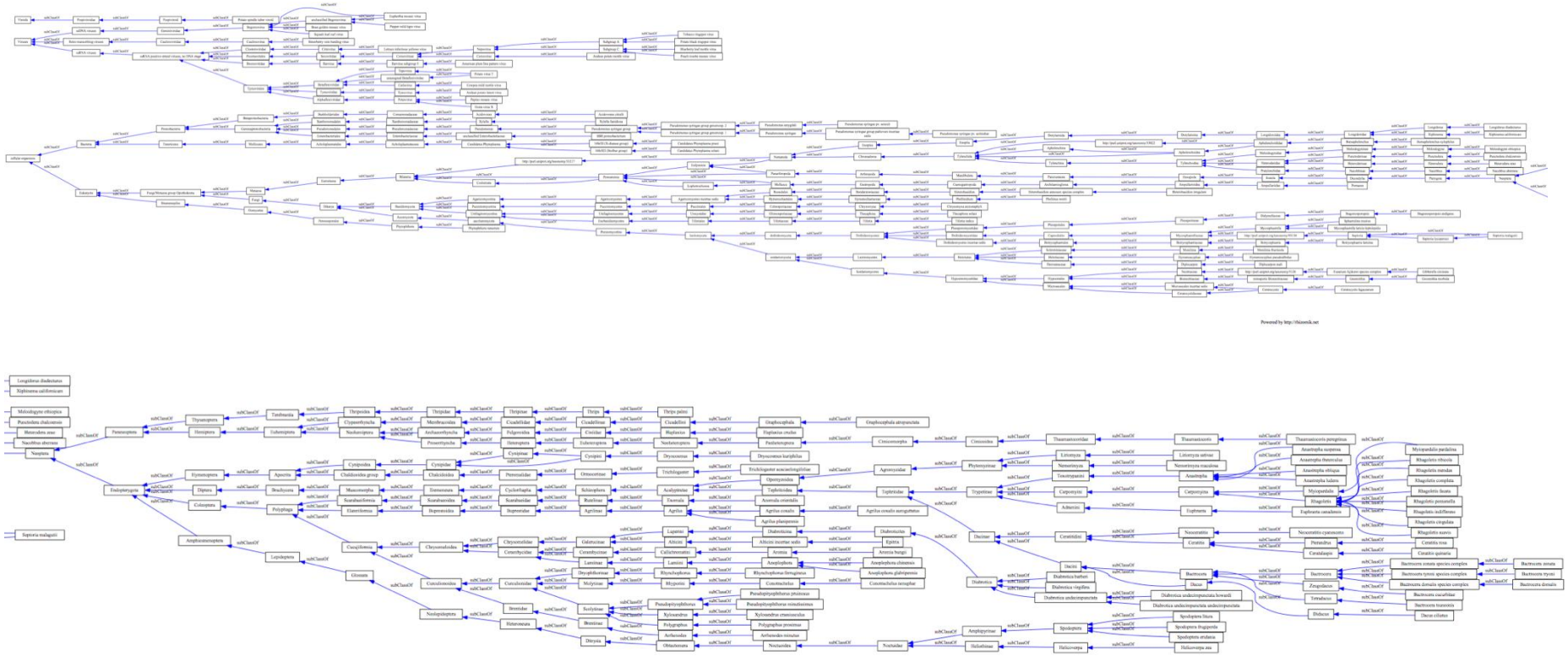
Parameter Name	Value-Metric translation	Evaluation process
Accessibility	Accessible: 3 Accessible by subscription: 2 Accessible by subscription without cost: 2 Restricted access: 1 Not accessible: 1	Physical conditions in which users can obtain data.
Relevance (A source relevant to find emerging/re-emerging plant risks)	3 (3 questions affirmative): 3 2 (2 questions affirmative): 2 1 (1 or no question affirmative): 1	Firstly, responds to if there is information about pest or plant health. And information about "interest pest" is searched". Besides, several questions are answered related to content: Information description (1), the following questions are not answered. But if, the resource is a "news" source, it is answered: Is there a causing agent identified in the information source analysed? (Y/N) Is there a specific crop identified in the information source analysed? (Y/N) Is there a region/country identified in the information source analysed? (Y/N)
Accuracy (about data described in content)	3 (3 questions affirmative): 3 2 (2 questions affirmative): 2 1 (1 or no question affirmative): 1	Accurate information provided: Following above answers as Yes, it is measured also: It is the risk identified? – Several detail levels could be found. Is a specific pest identified? Are there specific indicators of outbreaks (e.g. number of hectares affected or percentage of yield reduction)? If the first answers (relevance section) responses are NO, there is a poor accuracy level.
Edition (processed or raw data)	Raw: 1 Processed: 3	Maximum level (processed data), Minimum level (raw data) Raw (statistical information source or data tables, plots...) Processed (other kinds of sources)
Clarity	Metadata available: 3 Metadata not available: 1	Is there Metadata available (source title, description, etc.)?: Metadata available (there is title, description…) Metadata not available (there is just a title at most)
Comparability	Sufficient data: 3 Insufficient data: 1 Mixed comparability: 2	Sufficient data (maximum level); insufficient (minimum, e.g. only title of news); mixed (medium level)
Coherence		Is there a standard methodology

	Uses standard method or guideline: 3 None: 1	to present the information?
Authority	Yes: 3 No: 1	It is only possible to include 2 parameters, because a source has an authority identified or not. If there is one, it has de maximum scale level, if there is not information about an authority, it has the minimum scale level
Reputation	Yes: 3 No: 1	Reputation (only for journals and scientific communities) Official sources: they have a reputation implicit. It they are official (maximum level), if they are unknown (minimum level).
Timeliness	Sources should be monitored first to have measures for this indicator.	

Appendix C – IQ Metrics

Metadata Quality Perspective (2-point scale): presence or absence of 14 Dublin Core properties		**Content Quality Perspective (3-point scale)**	
Title		Relevance	
Creator		Accuracy	
Subject		Edition	
Description		Timeliness	
Publisher		Accessibility	
Contributor		Clarity	
Date		Comparability	
Type		Coherence	
Format		Authority	
Identifier		Reputation	
Source			
Coverage			
Language			
Rights			
Total Score:		Total Score:	
2x2 IQ Metrics (sum of both perspectives scores):			

Appendix D – Plant Health Threat Ontology Hierarchy



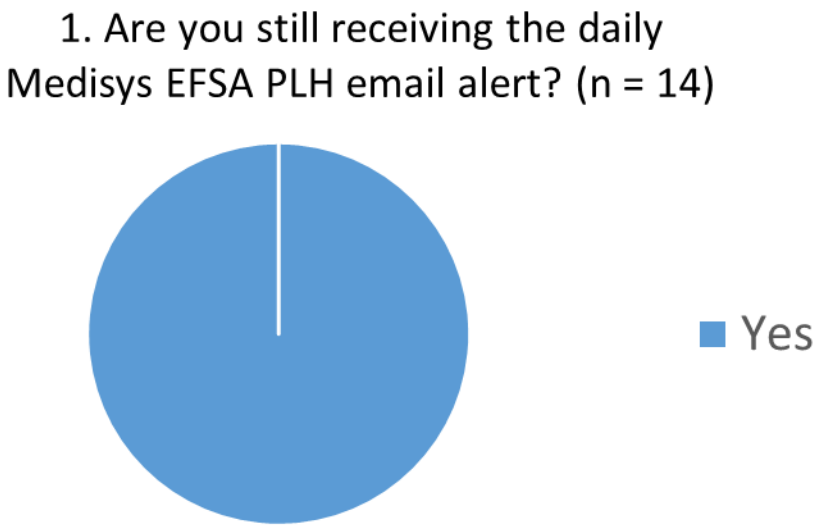
Annex A – Results of the EFSA Plant Health MediSys E-Mail Alert Survey

Marco Pautasso

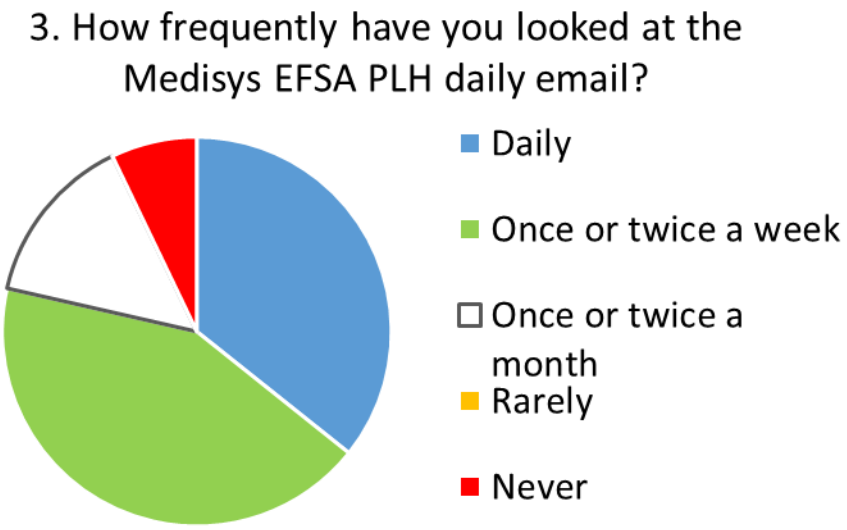
Animal and Plant Health Unit, European Food Safety Authority (EFSA)

Compiled on 21st June 2016

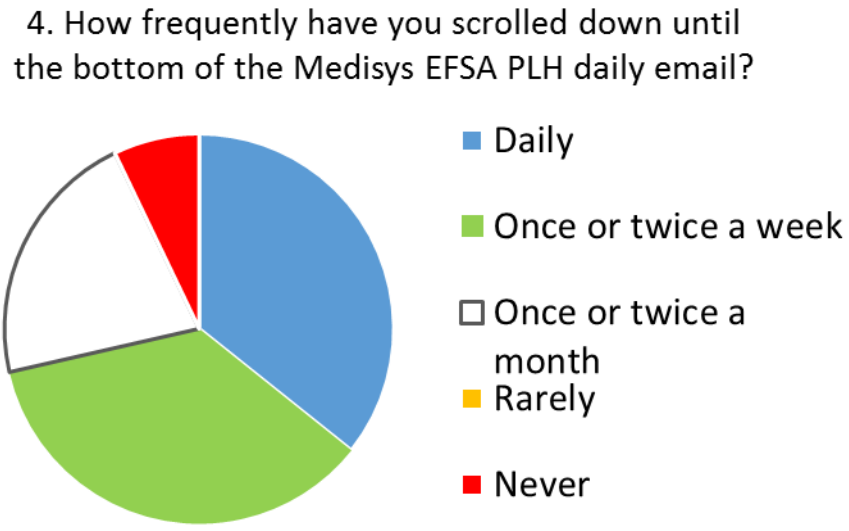
Reassuringly, all of the 14 people who answered the survey are still receiving the daily email alert. However, 29 people were originally registered to receive it, so about 50% did not complete the survey and we do not know whether they are still receiving the alert and what their view on it is. There could be a bias in the survey results as those who took the time to complete the survey might have more positive views on the alert than those who could not be bothered to do so.



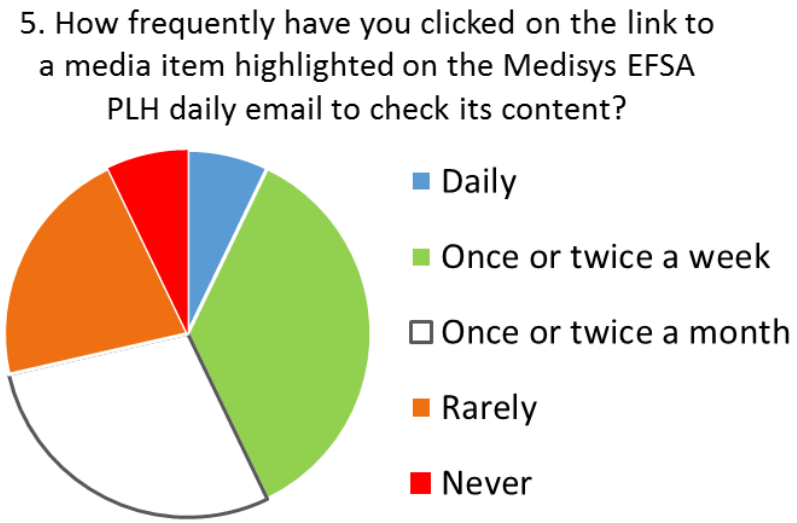
A majority of the respondents checked frequently (daily or once or twice a week) the alert. Only one respondent never did so during the several months of activity of the alert.



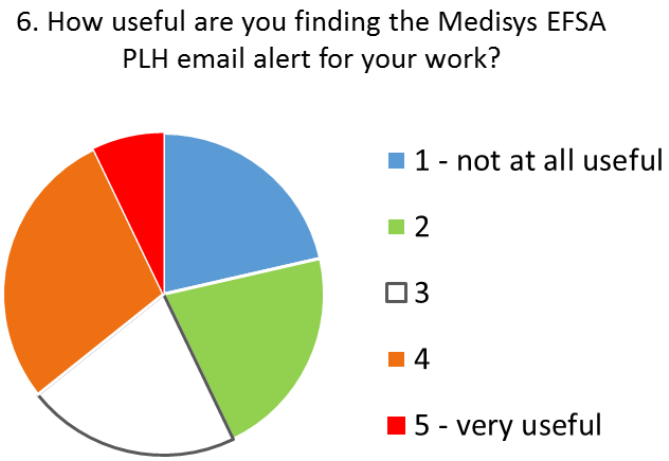
Most survey respondents also frequently scrolled down until the bottom of the email alert.



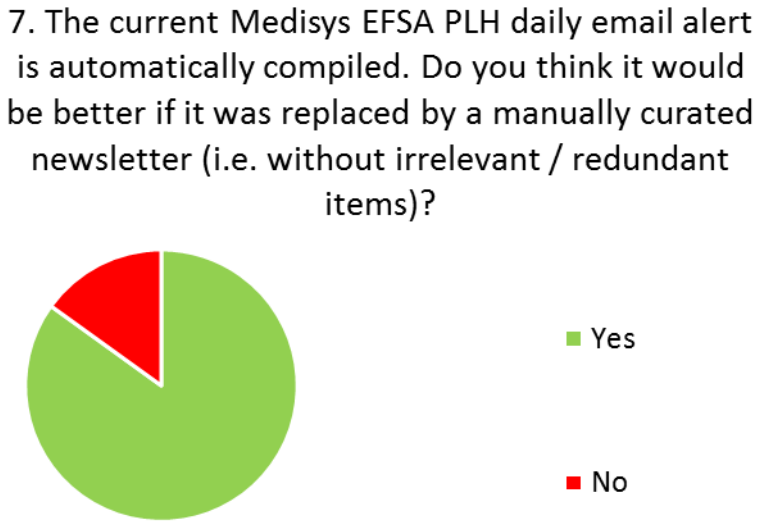
But there was a less frequent check of the media items highlighted in the email alert.



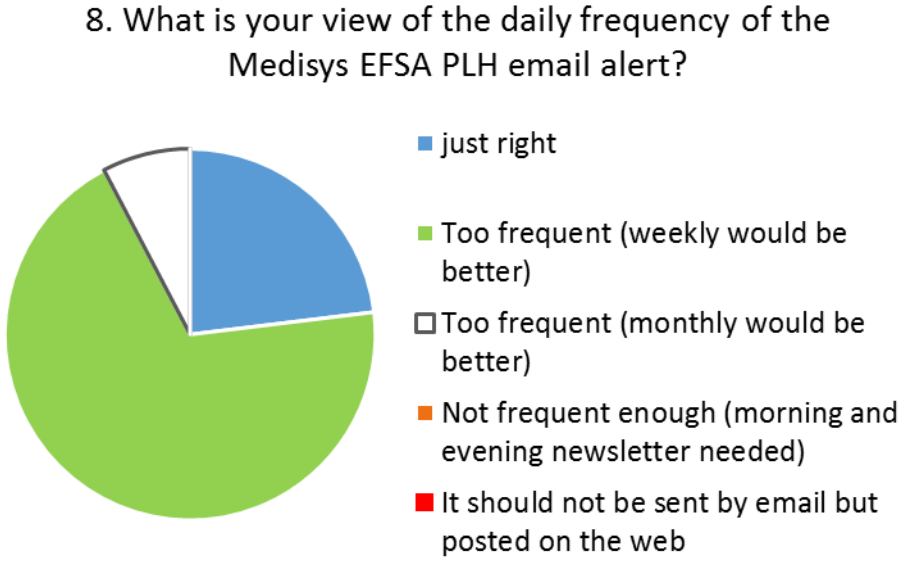
There is a balance between respondents finding the daily alert useful and those not finding it useful for their work.



Most respondents think that it would be better if the daily alert was replaced by a manually curated newsletter.

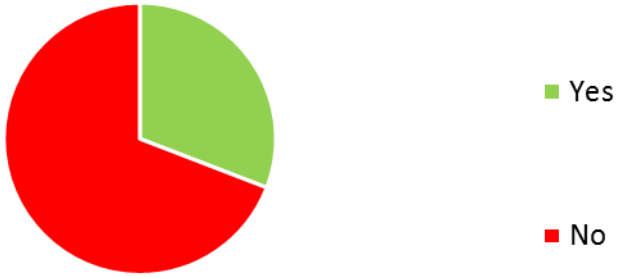


The majority of the respondents think that a weekly frequency would be better.



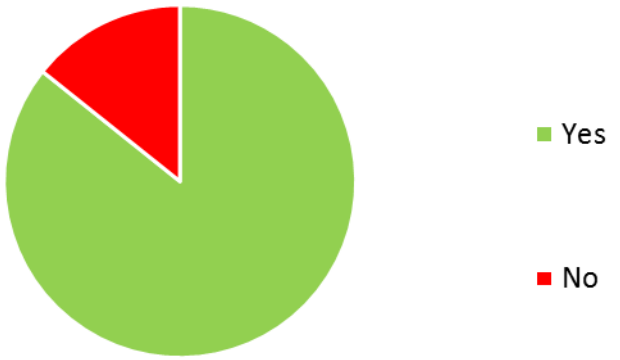
A minority of the respondents is making use of the MedIsys pages on individual plant pests.

11. Apart from the newsletter, are you making use of the Medisys EFSA PLH pages on individual pests or groups of pests?

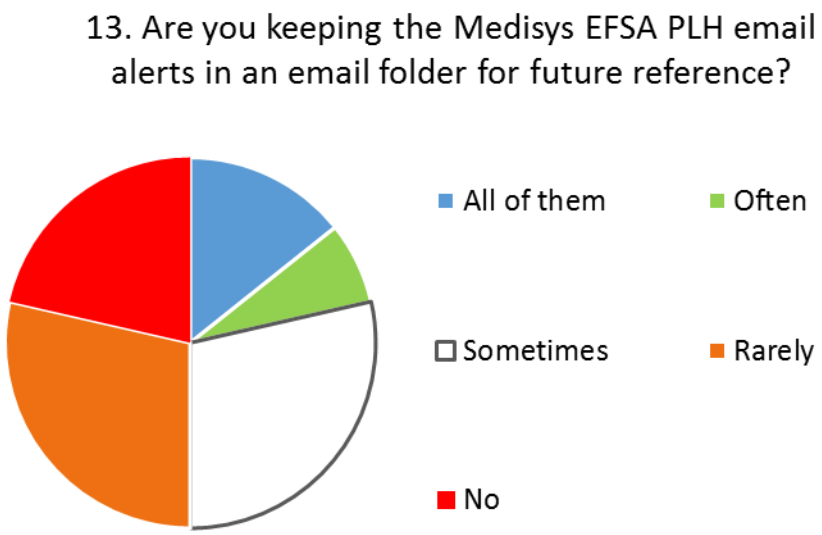


A majority of respondents would be happy to receive an additional newsletter about unknown emerging plant health threats.

12. Would you be happy to receive an additional email alert including media items from a search about unknown emerging plant health threats?



There is a balance between respondents keeping the alerts for future reference and those rarely or not doing so.



There was a balance among the respondents part of the EFSA PLH Panel, members of EFSA PLH Staff and other affiliations.